Universidade de Lisboa

Faculdade de Ciências

Departamento de Biologia Vegetal

Analysis of molecular diversity and *cis*-regulatory elements in the promoters of lignin genes in *Eucalyptus* species

Priscila Miriam Santos Pereira

Dissertação

Mestrado em Biologia Molecular e Genética

2014

Universidade de Lisboa

Faculdade de Ciências

Departamento de Biologia Vegetal

Analysis of molecular diversity and *cis*-regulatory elements in the promoters of lignin genes in *Eucalyptus* species

Priscila Miriam Santos Pereira

Dissertação

Mestrado em Biologia Molecular e Genética

Orientador interno: Professor Doutor Pedro Fevereiro

Orientador externo: Doutor Jorge Almiro B. C. Pinto Paiva

2014

“Science, my lad, is made up of mistakes, but they are mistakes which it is useful to make, because they lead little by little to the truth.”

Jules Verne

Agradecimentos

Gostava de mostrar o meu agradecimento a várias pessoas que das mais diversas formas foram essenciais para a realização deste trabalho.

Em primeiro lugar à Fundação para a Ciência e Tecnologia (FCT) e ao Projeto TreeForJoules (PKBBE/AGR-GPL/0001/2010), designado “Improving eucalypt and poplar wood properties for bioenergy”, financiado por fundos nacionais através da FCT/MCTES, no âmbito do qual foi desenvolvida esta dissertação de Mestrado.

Ao meu orientador Dr. Jorge Paiva, por guiar o desenvolvimento do trabalho, pela partilha do seu conhecimento e pelos seus úteis conselhos e correções.

Ao professor Pedro Fevereiro, orientador e diretor do grupo de Biotecnologia de Células Vegetais (BCV), por zelar pelo bem do laboratório, das pessoas e dos trabalhos. Por todas as sugestões e conselhos.

À Dra. Maria de Jesus Fernandes e ao Eng. Rui Pombo responsáveis pelo Departamento de Conservação da Natureza e Florestas de Lisboa e Vale do Tejo, pertencente ao Instituto da Conservação da Natureza e das Florestas, pela autorização da colheita de material vegetal na Mata Nacional do Escaroupim.

Aos colegas Victor Carocha e Filipe Cadete, pela preciosa ajuda no decorrer do trabalho, com esclarecimento de dúvidas e sugestões. Às colegas Susana Pera e Clara Graça, pela ajuda prática no decorrer do trabalho e pela amizade que já vem de trás. A todos os restantes colegas do grupo do BCV por me terem acolhido mais uma vez e proporcionado um bom ambiente de trabalho. Em especial, às colegas Olívia Costa, Maria Assunção e Rita Severino pelas palavras de ânimo e ajuda que disponibilizaram.

Um agradecimento especial às senhoras Fátima e Graça da Mata Nacional do Escaroupim, pela ajuda na recolha e identificação das amostras. Também pela simpatia e disponibilidade que sempre mostram.

Aos meus amigos, por serem uma parte essencial da minha vida e darem brilho a dias cinzentos. Pelo especial apoio, preocupação e sensibilidade a esta fase da minha vida gostava de destacar: Eliana, Filipa, tios Licas e Carlos, Ricardo, Adriana, Catarina, Vinicyus, Ruben, Rita, Diva e Eliseu. As vossas palavras e orações foram mesmo importantes.

À minha comunidade (Comunidade Cristã em Albergaria), a todos sem exceção, por serem família em todos os momentos. Por me lembrarem a razão da nossa esperança.

À minha família, o alicerce, o meu chão. Paizinhos não há palavras para descrever o que vocês são para mim. Sem vocês nada disto seria possível. Obrigada pela oportunidade, por nunca desistirem de mim mesmo quando falho. Tota, por seres meu maninho e isso bastar. Pelo teu carinho e conversas à distância.

Por último mas tão importante, a Deus. Por nunca falhar, por ser a minha força na fraqueza. Por ser a razão de uma vida com propósito.

Resumo

O género *Eucalyptus* tem grande importância comercial e ecológica. Os eucaliptos são as principais espécies produtoras de madeira usadas na indústria da celulose e pasta de papel. A quantidade e a composição em lenhina na madeira é uma das maiores barreiras para a qualidade da polpa. Vários estudos têm surgido para a identificação dos mecanismos de transcrição e pós-tradução envolvidos no controlo da lenhificação. Apesar disso, são muitas as lacunas na compreensão desses mecanismos, nomeadamente no género *Eucalyptus*.

A regulação da transcrição é determinante na regulação da expressão génica. A região promotora contém elementos *cis*-regulatórios que se ligam a fatores de transcrição específicos, dirigindo a expressão espacial e temporal da transcrição. O conhecimento da composição dos promotores em elementos *cis*-regulatórios é essencial para a compreensão dos mecanismos inerentes à regulação da transcrição. Nos últimos anos têm surgido várias ferramentas bioinformáticas para a predição destes elementos nos promotores. Este tipo de metodologia é desenvolvido, maioritariamente, por comparação de regiões reguladoras ortólogas, de um mesmo gene em várias espécies com diferentes graus de divergência e polimorfismo. Desta forma tem sido possível, ao perceber quais os motivos com maior probabilidade de possuírem função biológica, economizar tempo e recursos no decorrer do projeto experimental.

Os fatores de transcrição (FT) da família MYB ligam-se a elementos *cis*-regulatórios específicos nos promotores de vários genes que codificam para enzimas envolvidas nas vias biossintéticas dos fenilpropanóides e lenhina. A existência comum destes elementos nos promotores de vários genes da via da lenhina tem suportado a hipótese de que estes são regulados de forma coordenada a nível transcricional. Existem outros FT, que em cooperação com os FT da família MYB, são responsáveis pela regulação coordenada dos genes da via biossintética da lenhina. A melhor compreensão desta via passa pela descoberta de novos elementos *cis*-regulatórios, tendo em conta a evolução e função de famílias de genes da biossíntese de lignina em todo o género *Eucalyptus*.

Nesta dissertação é descrita a análise da região promotora de quatro genes envolvidos na via biossintética da lenhina, *PAL9*, *F5H2*, *CSE* e *CAD3*, isolados a partir de nove espécies do género *Eucalyptus*. A análise foi realizada em termos de diversidade nucleotídica, composição e conservação em elementos *cis*-regulatórios.

Dos 30 pares de *primers* desenhados para amplificação por PCR da região promotora de cada gene pertencente à via biossintética da lenhina em eucalipto, foram escolhidos 11 para amplificação da região promotora de 62 espécies do género *Eucalyptus* e quatro do género *Corymbia*. Dez dos 11 pares de *primers* amplificaram com sucesso em mais de metade das espécies analisadas, sendo que o sucesso total na amplificação foi de 73%. Estes resultados confirmaram a existência de um grau de conservação considerável entre os promotores ortólogos. As espécies com menor sucesso na amplificação foram as espécies pertencentes ao género *Corymbia* e aquelas pertencentes ao subgénero *Eucalyptus* (género *Eucalyptus*). Estes resultados eram espectáveis uma vez que estas espécies são as mais afastadas filogeneticamente da espécie *E. grandis*, para o qual os *primers* foram desenhados.

Com base na hipótese de que os promotores contêm regiões específicas mais conservadas, garantindo a manutenção e funcionalidade de elementos *cis*-regulatórios, foi realizada uma abordagem genómica comparativa ao nível da espécie. Para isso, as regiões promotoras dos genes *PAL9*, *F5H2*, *CSE* e *CAD3* foram amplificadas por PCR para um indivíduo de cada uma de 9 espécies do género *Eucalyptus*, 7 delas pertencentes ao subgénero *Symphyomyrtus.* Posteriormente, estas regiões promotoras foram diretamente sequenciadas, e em alguns casos primeiramente clonadas e só depois sequenciadas. Foram seguidamente alinhadas para uma análise mais detalhada em termos de diversidade nucleotídica (π) e para identificação e mapeamento de elementos *cis*-regulatórios. De uma forma geral a diversidade nucleotídica foi elevada. Os valores de π variaram de 0.026, para as regiões promotoras *PAL9* a 0.078 para as regiões promotoras *CAD3.* O facto da enzima fenilalanina amónia-liase (*PAL*) ser essencial para a síntese de todos os fenilpropanóides, catalisando o primeiro passo dessa via biossintética, pode ser a razão da manutenção deste promotor ao longo do processo de divergência e especiação.

Apesar dos elevados valores de π, foram identificadas regiões específicas, em todos os grupos de promotores, com valores de diversidade nucleotídica espectáveis para regiões codificantes conservadas (π inferior a 0.02), podendo significar a existência de restrições funcionais para essas regiões.

O conjunto de elementos *cis*-regulatórios putativos mapeados ao longo das regiões promotoras analisadas foi obtido por uma seleção prévia através de duas estratégias. A primeira foi a identificação de elementos *cis*-regulatórios por homologia com elementos *cis*-regulatórios descritos na base de dados PLACE. Essa identificação foi

realizada nas regiões promotoras (1000pb anteriores ao ATG) de todos os genes pertencentes à via biossintética da lenhina em eucalipto e dos respetivos ortólogos nas espécies *A. thaliana*, *P. trichocarpa* and *V. vinífera,* consistindo num total de 75 sequências. De um total de 219 elementos identificados, foram selecionados 32 elementos representativos do conjunto de dados, sendo mais provavelmente envolvidos na regulação génica da lenhina. A segunda estratégia utilizada consistiu na identificação *in silico* de motivos sobrerepresentados no conjunto das mesmas 75 sequências, utilizando para essa deteção três programas, MEME, MotifSampler e oligo-analysis (RSAT). A ferramenta STAMP foi utilizada para agrupar os 77 motivos, identificados pelo conjunto dos três programas, por similaridade e redundância. Isto permitiu obter um conjunto de cinco novos elementos-*cis* putativos que não seriam normalmente identificados por pesquisas em bases de dados. Três deles foram identificados por mais do que um programa, aumentando a confiabilidade e precisão da predição.

O mapeamento do conjunto de elementos *cis*-regulatórios previamente selecionados, nos quatro grupos de promotores, permitiu verificar a existência de um padrão de conservação nas espécies de eucalipto.

Ao analisar o efeito potencial da evolução das sequências de DNA nas ocorrências de elementos *cis*-regulatórios identificados em regiões de baixa diversidade nucleotídica, nos quatro grupos de promotores, verificou-se que mais de metade dessas ocorrências de elementos-*cis* (52%) eram conservadas nas sequências promotoras das nove espécies em análise. Ainda assim, foram identificados alguns casos, a maioria em *E. tereticornis* (56%), onde alterações nucleotídicas causaram a perda de determinado elemento-*cis* putativo, nas espécies onde essas alterações ocorreram. As diferenças encontradas na região promotora de mais do que um gene, na espécie *E. tereticornis*, poderão ter implicações na regulação transcricional dos genes da lenhina nesta espécie.

Os elementos CRPE31, GT1CONSENSUS, SEF4MOTIFGM7S, PYRIMIDINEBOXOSRAMY1A e o motivo sobrerepresentado LRPE2, identificado *de novo* neste estudo, foram identificados com múltiplas ocorrências em todos os grupos de promotores ortólogos. Indicando a sua eventual importância na regulação de genes com expressão sincronizada ou no controlo dos níveis de mRNA. Para além disso, todos estes elementos foram identificados, nas regiões promotoras de um ou mais genes, em regiões com valores de π inferiores a 0.02 podendo ter potencialmente maiores restrições funcionais. Adicionalmente, todos estes elementos, exceto o

PYRIMIDINEBOXOSRAMY1A, parecem estar posicionalmente conservados nas regiões promotoras de genes diferentes. Esta informação pode ser relevante pois a posição relativa dos elementos tem influência nos perfis de transcrição dos genes. Todos os elementos-*cis* necessitam de validação experimental. Ainda assim, todas estas evidências em consonância com as funções biológicas dos vários elementos descritas em literatura permitiram perceber de forma mais concreta quais as funções putativas destes elementos nos promotores dos genes da lenhina em *Eucalyptus*.

Em conclusão, este trabalho permitiu compreender novos aspetos da diversidade e evolução dos promotores de genes da lenhina em *Eucalyptus.* Além disso, foram revelados elementos *cis*-regulatórios que podem ser responsáveis por aumentar a atividade dos promotores ou pelo controlo temporal e/ou espacial da expressão de vários genes no decorrer da biossíntese da lenhina em *Eucalyptus*. Esta informação será capaz de acelerar a caracterização funcional de novos elementos *cis*-regulatórios e dos respetivos FT, sendo um passo fundamental na compreensão da regulação transcricional da via biossintética da lenhina e da sua heterogeneidade no género *Eucalyptus*.

Abstract

The genus *Eucalyptus* has great commercial and ecological importance, being the principal hardwood species used for pulp and paper manufacturing. The fact that the quantity and composition of lignin in wood has implications in pulping quality and bioethanol production has made lignin biosynthesis an extensively studied pathway. However, more information regarding diversity and transcriptional regulation of the genes involved in the lignin regulation biopolymer is needed.

In this work a comparative analysis of orthologous phenylpropanoid-lignin genes promoters from four angiosperm genera *Eucalyptus*, *Arabidopsis*, *Populus* and *Vitis*, was performed to identify overrepresented conserved and novel motifs being putative *cis*-regulatory sequences (CREs). Comparative analysis identified 33 conserved sequence motifs, consisting of known and novel CREs, significantly overrepresented in those promoter sequences. Subsequently, were analyzed patterns of species-level nucleotide diversity and the distribution and composition in putative *cis*-regulatory elements in the promoters of four phenylpropanoid-lignin genes from different *Eucalyptus* tree species using *in silico* tools. Species-level nucleotide diversity (π) varied from 0.026 to 0.078, and the different lignin promoters studied showed different patterns of sequence conservation. All the promoters of the studied *Eucalyptus* lignin genes contained highly conserved regions at the species level. Some putative CREs were identified in all the promoters with multiple occurrences and positioned in specific regions with low nucleotide diversity. This suggested functional restrictions on such regions and biological function for those specific motifs.

These findings provide new information capable to accelerate the functional characterization of CREs and their interacting TFs responsible for driving high or specific activity of corresponding promoters in lignin biosynthesis in *Eucalyptus*. Thus, this work contributes to a step forward in the understanding of the mechanisms underlying the regulation of lignin heterogeneity in *Eucalyptus*. This knowledge, in turn, will benefit *Eucalyptus* breeding programs or enable the alteration of lignin composition, increasing its degradability in industrial processes.

Key words: *cis*-regulatory elements, promoter evolution, lignin, phenylpropanoid biosynthetic pathway

Contents

List of figures and tables

1 INTRODUCTION

Natives from Australia, eucalypts, are one of the most valuable and widely planted hardwoods with more than 20 million hectares planted worldwide [1]. *Eucalyptus* is an angiosperm dicotyledonous genus, belonging to the *Myrtaceae* family [2, 3]. Currently, about 800 eucalypt species can be recognized [4-6]. This genus has a huge commercial and ecological importance due to their interesting properties including fast growth, superior wood and fiber properties, outstanding diversity and adaptability [1, 2]. According to the United Nations Food and Agriculture Organization, eucalypts are the principal hardwood used for pulp and paper manufacturing [7]. The sector of pulp and paper is of great importance for the economy of Portugal. In this country forests occupy about 35% of the national territory and *Eucalyptus spp*. are the first most abundant species occupying about 812.0 hectares [8].

Pulp and paper manufacturing is highly dependent of the wood quality, mainly of the composition and structure (physical and chemical) of wood traits [9]. A major barrier for pulping quality and bioethanol production is the quantity and composition of lignin in wood [10, 11]. Lignin is a complex polyphenolic heteropolymer of high carbon content mainly present within the secondary cell wall of xylem elements, and represents the second most abundant terrestrial biopolymer, after cellulose [2, 11]. These aspects have made lignin biosynthesis the most studied pathway involved in wood formation, with great interest in the knowledge and comprehension of its structure, regulation and function [10, 11].

Lignins are generated by the oxidative polymerization of the three monomers of p-hydroxycinnamoyl alcohols through the general phenylpropanoid pathway and the monolignol-specific pathway. These monomers (p-coumaryl, coniferyl and sinapyl alcohols) known as monolignols give rise, upon incorporation into the lignin polymer, to the p-hydroxyphenyl (H), guaiacyl (G) and syringyl (S) units (Fig. 1) [10-12]. The monolignol biosynthetic pathway and the enzymes involved are represented in Fig. 1. Generally, in many plants, these enzymes are found as multiple isoforms encoded by different genes belonging to multigenic families [9, 11, 13]. Great progresses in the understanding of wood biosynthesis in forest trees, including *Eucalyptus species*, have been achieved. Despite this, more information regarding the genes involved in the regulation of lignin polymerization is needed [2, 13-15].

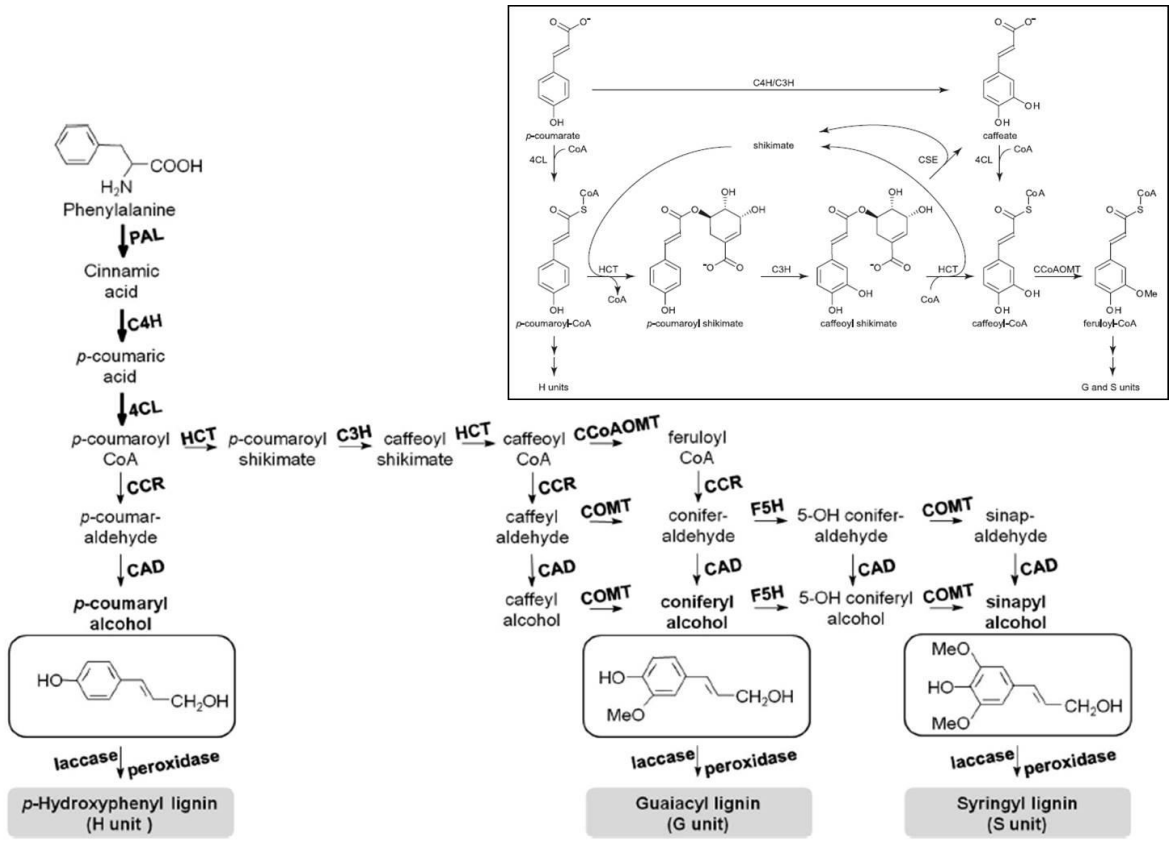

Fig. 1: The monolignol biosynthetic pathway. Enzymes from the general phenylpropanoid pathway, **PAL** (phenylalanine ammonia-lyase), **C4H** (cinnamate-4-hydroxylase) and **4CL** (4-coumarate:CoA ligase), are represented by thick arrows. Enzymes from the monolignol-specific pathway, **C3H** (p-coumarate 3-hydroxylase), **F5H** (ferulate 5-hydroxylase), **CCoAOMT** (caffeoyl-CoA O-methyltransferase), **COMT** (acid/5-hydroxyferulic acid O-methyltransferase), **CCR** (cinnamoyl CoA reductase), **HCT** (hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyl transferase) and **CAD** (cinnamyl alcohol dehydrogenase) are represented by thin arrows. Monolignols are indicated in bold above their corresponding molecules in boxes. The box at the top represents the lignin biosynthetic pathway incorporating the **CSE**-dependent reaction established by Vanholme et al. (2013) (Adapted from [11, 16]).

Recently, a "molecular *toolbox*" for lignin biosynthesis in *Eucalyptus*, was identified in *Eucalyptus grandis* [13]. The combination of comparative phylogenetic and expression analysis suggested a *Eucalyptus* core lignin *toolbox* featuring 17 members belonging to eleven multigene families. In addition, five potentially novel genes not reported yet in any other species were also identified. These new genes were founded highly or preferentially expressed in forming xylem, and may play important roles in the lignin/phenylpropanoid pathway [13]. This core lignification gene set containing known and potentially novel candidate genes for future functional studies in *Eucalyptus* provided the theoretical foundation for this work.

The knowledge about the genetic architecture and diversity within *Eucalyptus* species is of great importance when regarding the study of genes putatively controlling economic or adaptive traits [14]. Research at the level of evolution and diversity of regulatory networks underlying wood development has been facilitated by the high

nucleotide diversity found in the *Eucalyptus*, and the recently completed *E. grandis* reference genome sequence [1, 11, 17, 18]. These studies, often by means of comparative analyses of orthologous gene regions across multiple species have made possible to make inferences about the direction of wood biosynthesis [14].

In eukaryotes, there are several mechanisms regulating gene expression. Nevertheless, the initiation of transcription seems to be the primary determinant of control of gene expression [19, 20]. The 5' gene flanking-region, called promoter region, contains several regulatory sequences which ensure the spatial and temporal regulation of gene transcription. It contains binding sites allowing specific interactions with transcription factors and RNA polymerase [21]. Transcription factors (TFs) are proteins that bind to short DNA sequences (5-20bp) influencing the status of the pre-initiation complex and in turn, direct the expression of target genes [19, 22]. Short DNA signals to which the TFs interact are called *cis*-regulatory elements (CREs) [19, 23]. Specific associations between TFs and CREs allows to a combinatorial effects at the level of individual promoters, enabling the achievement of meaningful transcription rates and determining transcriptional responses at internal and external stimuli [19, 24].

Understanding about *cis*-regulatory elements (CREs) plurality, attendance, relative position, orientation and composition of sequence in promoters, contributes to provide insights into the mechanisms underlying the transcriptional regulation of genes [20]. Combining CREs data with co-expression data and gene ontology annotation is possible to link many genes and motifs to specific biological functions. Ultimately, this understanding might allow temporal and spatial modulation of gene expression [20, 24].

Traditionally, many studies were performed through the challenging construction of promoter deletion series and analysis in transgenic cell lines or organisms in view of the elucidation of gene networks [24-30]. However, this approach is laborious, time consuming and limited, since only a small fraction of bases within intergenic regions serve as TFs binding sites.

In the last few years, several bioinformatic strategies have been used to predict potential regulatory regions within promoters. All the predictions of novel interactions between TFs and CREs need to be experimentally validated. However, *in silico* approaches allow conducting properly the experimental design, saving time and resources, by choosing those motif predictions whose experimental validation is worthwhile [24, 31, 32]. Numerous tools have become accessible for prediction of motif as good candidates for regulatory elements [22, 23, 31, 33], such as MEME [34],

MotifSampler [35] and RSA Tools [21]. Bioinformatic methods can also be used to scan sequences for matches against some databases of motifs previously reported as *cis*-acting regulatory DNA elements, such as the plant database PLACE [36, 37].

The comparison of orthologous regulatory regions of a single gene from several species of variable degrees of divergence and polymorphism has been the most popular and successful approach to deduce putative regulatory elements, notably in *in silico* analysis [15, 33]. This approach implicitly infers that the functional regions of DNA are under evolutionary pressure and stay conserved in so far as they are functionally constrained [24, 33]. Computational algorithms for *in silico* detection of CREs take into account their conservation in orthologous promoters from different species and genera, due to the maintenance of regulatory networks. Comparative analysis of the promoters of co-expressed or co-regulated genes presents another alternative to identify shared functional CREs, and those algorithms also take that feature into account [15, 17, 22, 38, 39]. The accurate identification of regulatory motifs embedded in promoter regions of genes is not trivial and in plants this becomes a particularly challenging task 31]. Although challenging, this *in silico* approach for CREs detection already led to the identification of many known transcription factor binding motifs and new putative regulatory motifs with unknown function in promoters of several genes in a number of plant species including *Populus* and *Eucalyptus* [15, 19, 39-41].

Over time, a number of studies emerged centered in the identification of the transcriptional and posttranslational mechanisms involved in the temporal and spatial control of lignification. Global transcript profiling analysis and functional characterization was mainly used to achieve this end [10-12]. Lately, research in regulation of lignification has focused on the promoters of the genes of the phenylpropanoid metabolism and lignin branch pathway and in the TFs that can interact with them. Those results support that the coordination of the expression of genes encoding the enzymes of the lignin biosynthetic pathway is regulated at the transcriptional level [10, 12, 42].

Several studies showed that a number of TFs families, especially plant MYB proteins, participate notably in the regulation of the phenylpropanoid and lignin specific branch biosynthetic pathways. Those TFs were found to bind to a specific type of CREs, AC - rich *cis*-acting elements, driving gene expression in forming xylem tissues in several plants [10, 11, 28-30, 43]. AC-rich *cis*-acting elements, more recently entitled MYB *cis*-elements, were identified as necessary for xylem expression for the first time in a bean *PAL2* promoter [27]. Later, the same authors demonstrate that those elements were

involved in conferring tissue-specific expression also in *PAL3* promoter [28]. Several subsequent studies, in different plant species including eucalypts, identified their presence and conservation in the promoters of other genes encoding enzymes implicated in lignin biosynthesis, like *C4H*, *4CL*, *HCT*, *C3H*, *CCoAOMT*, *CAD* and *CCR* [10-12, 25, 26, 29, 42-44]. The common existence of CREs of MYB type in several lignin specific gene promoters has hinted the existence of a mechanism by which different steps of phenylpropanoid metabolism are coordinately regulated [12, 28, 29, 43, 44].

Although the indispensable presence of MYB TFs other transcription factors are coming to light [12, 29]. Rahantamalala et al. [43] detected other putative CREs possibly serving as binding sites to other TFs which in turn, cooperating with MYB TFs, may function to lignin biosynthetic gene coordinated regulation [43].

Despite all the efforts and advances made so far, many aspects regarding the mechanisms involved in transcriptional control of lignification during the process of secondary cell wall formation remain unsolved [2, 11]. Further research is required in order to identify other TFs that bind to functional validated CREs, taking into account the evolution and function of lignin biosynthesis gene families across the genus *Eucalyptus* [43]. In the last instance the integration of all this knowledge will allow the application of genomic technologies, benefiting *Eucalyptus* breeding programs or enabling the alteration of lignin composition, increasing its degradability in industrial processes [2, 10, 11].

2 OBJECTIVES

This work aims at the *in silico* characterization of upstream promoter sequences of co-expressed lignin genes in evolutionary distinct *Eucalyptus* tree species. The specific objectives were: a) to quantify and assess patterns of species-level nucleotide diversity in the promoters of *Eucalyptus* species; b) to identify putative novel CREs by a comparative analysis of *Eucalyptus* lignin genes promoters with orthologous promoters from *Arabidopsis*, *Populus* and *Vitis;* c) to identify putative *cis*-element occurrences within the *Eucalyptus* promoters and their conservation within these species; d) to provide information capable of accelerate the functional characterization of CREs and their interacting TFs; e) to serve as a step forward in the understanding of the transcriptional regulation of the lignin biosynthetic pathway in the genus *Eucalyptus*.

3 MATERIAL AND METHODS

3.1 Plant Material and Nucleic Acids Extraction

The plant material used in this study was collected in Eucalyptus Arboretum at Escaroupim National Forest (Escaroupim, Portugal) in October 9 and 11 and November 11, 2013. Leaves were sampled on individual trees from single species collection plots [45]. The sampling resulted in a total of 62 species (Table S1). All plant material was lyophilized using a freeze dryer and properly stored until further use. Exceptionally, *E. grandis* young leaves of clonally *in vitro* propagated plants were sampled.

The DNA extraction was performed according to the protocol adapted from Gemas [46]. Approximately 50-100mg of lyophilized leaf tissue discs, from a single tree of each of 62 species were used for DNA extraction. DNA integrity was checked by running the DNA samples in a 0.8% (w/v) agarose gel, in 0.5X Tris-Borate-EDTA buffer (TBE), and stained SYBR Safe (Invitrogen®, Carlsbad, CA, USA), along to the molecular marker λ phage DNA (Invitrogen®, Carlsbad, CA, USA). Bands were visualized in a Gel Doc-1000 UV (Bio-Rad® Laboratories, Inc.) image acquisition system. Quantification and quality (A260/280 and A260/230 ratios) of nucleic acids was determined using the NanoDrop ND-1000 Spectrophotometer (Thermo ScientificTM, Wilmington, Delaware, USA).

3.2 Selection of promoter region

Eucalyptus lignification genes promoter region sequences:

The 1000bp upstream the first start codon (ATG) of each gene from the *Eucalyptus* core lignin *toolbox* [13] (Table S4) were chosen to represent promoter region. A sequence containing the 1000bp promoter regions and the 200-300bp after the first start codon were extracted from the current annotation set version 1.1 of the *E. grandis* genome stored in JGI/CIG-Phytozome v9.1 using the PhytoExtract [47].

Arabiopsis, *Populus* and *Vitis* lignification genes promoter region sequences:

Orthologous promoter sequences, of the gene members belonging to the *Eucalyptus* core lignin *toolbox*, of *A. thaliana*, *P. trichocarpa* and *V. vinifera*, three angiosperm genera were retrieved from the available versions of the genomes stored in Phytozome v.10 [63]. For orthologs selection the tool Keyword search was used, using *E. grandis* genes as target. Then the respective protein homologs of each of the three species with higher score and percentage of similarity were selected (Table S4). The

1000bp upstream flanking-coding region (Transcript) sequence from each orthologous gene was retrieved using the sequence retrieval tool from BioMart [48]. The promoter datasets were compiled resulting in a total of 75 sequences representing the members of the *Eucalyptus* core lignin *toolbox* and their orthologous promoter regions in the other three genera.

3.3 Primer design

Primer 3 Plus (http://primer3plus.com/) was used to design new primer pairs that targeted the maximum of the promoter region sequences, with the following options: i) annealing temperature of 60ºC and G/C content 40-60%; ii) primer size between 18 to 22bp, and expected product size range of 851-1000bp; iii) Max Self Complementarity and Max Pair Complementarity of 0 whenever possible. The forward primers were designed to match the 5' promoter region sequence end and the reverse primers to match the first nucleotides of the transcript sequence. For all genes but *EgrHCT5* was possible to design specific set of primers (Table S2). For gene *EgrHCT5* it was no possible design a set of primes due to the high percentage of N bases in the promoter sequence.

3.4 DNA Amplification

The PCR amplification was performed using the 50 ng DNA of each species sample (Table S1) and the primers described in Table S2. Water was used as negative control. All amplifications were held in Thermal Cycler C-1000 (Bio-Rad® Laboratories, Inc.), in a total volume of 20µL using 10ng of DNA, 5X Colorless GoTaq® buffer, 1.5mM MgCl2, 0.2mM of each dNTPs, 0.4µL of primer forward (10µM), 0.4µL of primer reverse (10µM), and 1U of GoTaq® DNA polymerase (Promega, Madison, WI, USA). Conditions of the PCR amplification were as follows: denaturation cycle at 94ºC for 3 min, then 40 cycles at 94ºC for 15 seconds, annealing at 60ºC or 65°C for 30 seconds, extension for 72ºC for 1 min, and a final extension at 72ºC for 10 min. The evaluation of the bands of the PCR products was performed using the 2% (w/v) agarose 0.5X TBE gel and the molecular marker 1 Kb Plus DNA Ladder (Invitrogen®, Carlsbad, CA, USA). The bands were visualized in a Gel Doc-1000 UV image acquisition system (Bio-Rad® Laboratories, Inc.).

3.5 Sequencing of promoter regions

In order to prevent unincorporated components and undesired by-products, which interfere with direct DNA sequencing, PCR products were purified with NZYGelpure kit (NZYTech), according to the manufacturer's instructions. All the purified products were eluted in 20µl or 30µl of ultrapure water and quantified before sequencing. The PCR products were sequenced at STAB Vida (STAB Vida, Lisboa, Portugal) using reverse primers in order to maximize the promoter region sequence.

For those genes/samples combinations in which the sequences presented multiple overlapping sequence peaks, the sequence was discarded, the corresponding PCR reactions repeated, and the PCR products were cloned using the TOPO® TA Cloning® Kit (Invitrogen®, Carlsbad, CA, USA). The transformed colonies were identified, handpicked and cultured overnight in LB containing 50µg/mL ampicillin. Cloned fragments were checked by colony PCR screening using universal primers: M13-20 forward (5´-GTAAAACGACGGCCAG-3´) and M13 reverse (5´-CAGGAAACAGCTATGAC-3´). The correct size PCR products were then purified and the amplified product corresponding to one colony was random selected and sent for sequencing, using one of the universal primers: M13-20 forward and M13.

3.6 DNA sequence analysis

Chromatograms and quality chart of each sequence were analyzed to check sequence quality (minimum phred 20). Cloned fragment sequences were scanned with Vecscreen [49] to identify and remove any segments of vector origin. When present in the sequences, the downstream ATG sequences were removed for further analysis. Sequences were aligned using ClustalW Multiple alignment function [50] of the BioEdit Sequence Alignment Editor version 7.2.5 [51] and alignments were corrected by manual inspection.

DnaSP (DNA Sequence Polymorphism) version 5.10.01 software [52] was used to calculate several measures of genetic polymorphism including the nucleotide diversity, Pi (π; [53]). For all nucleotide diversity calculations were used the associated algorithm Nucleotide diversity (gaps/missing data). This option is used in the events that there are gaps in the sequence data. Otherwise the *indels* present in the sequences would be ignored. Nucleotide diversity distributions along each promoter region sequence were graphically represented using the sliding window method, calculating π on a moving window of 50bp with a step size of 10bp.

3.7 Putative *cis*-regulatory elements *in silico* prediction and selection

Two strategies were used to select putative *cis*-regulatory elements (CREs) to be mapped into the *Eucalyptus* sequences: i) identification of known CREs and ii) identification of novel putative CREs.

First, CREs were chosen by searching known motifs in the sequences of promoter region data set (3.2 Material and Methods). Each of the promoter regions were individually scanned using the plant-specific database PLACE [37]. *Cis*-elements with motif lengths greater than five base pairs were selected if they were present in promoters of genes of all the 11 multigene families composing the *Eucalyptus* lignin *toolbox* and present in at least 50% of the scanned sequences. In addition, five secondary cell wall-associated elements previously reported in literature were also retrieved for subsequent *cis*-elements mapping, as follows: CRPE31, CRPE28, CRPE26, CRPE25 [17] and PALBOXPPC [54].

A second group of putative CREs motifs were predicted based on their overrepresentation in the sequences data set of promoter regions (3.2 Material and Methods). Three software programs were used for motifs discovery: a) MotifSampler [35, 55]; b) RSA tool oligo-analysis [21, 56]; c) MEME [34, 57]. An element was considered overrepresented when their frequency in the 75 examined promoters was above the average level of the *Arabidopsis* genome, the background model available and selected in the three software programs. For the purposes of this work the default settings of the programs were used, with exception of those specified or below:

a) MotifSampler [35]:

The output parameter - r (number of times the motif detection algorithm must be repeated) was set to 50. Was used the *Arabidopsis* background model of order 4 provided by the software package, giving more reliability to high LL scores obtained.

b) RSA tool oligo-analysis [21]:

The oligomer lengths selected were of size 7 and 8, excluding the ones of size 6. The conversion of assembled patterns to matrices was defined to a maximum of 10 matrices.

c) MEME (Multiple Em for Motif Elicitation) [34]:

The width of the motif was defined with a minimum of 6bp and a maximum of 8bp. The maximum number of motifs to find was defined as 4 motifs (rather than the default 3). Taking into account the distribution of motif occurrences among the sequences it was performed a run with the option - Zero or one occurrence of a single motif per sequence and a second run with the option - Any number of repetitions per sequence.

STAMP [58, 59] was used to compare and to hierarchically cluster, by motif similarity and redundancy, the outputs of the three programs. Motifs were inputted into the software as position-specific scoring matrix (PSSM) allowing a more reliable comparison taking into account the "background" probability of different letters. The clustering resulted from a joint analysis of the proximity of the motifs in the UPGMA tree (output of STAMP) and of the observation of the correspondent most similar CREs in the PLACE database for each motif. The motifs of each defined group were aligned and a consensus sequence was generated. The motifs were entitled lignin-related promoter elements (LRPE) and numbered sequentially.

3.8 *Cis*-element mapping

RSA tool Pattern Matching [21, 56] was used to map and count occurrences of the selected CREs in the promoter sequences of each of the selected *Eucalyptus* species. Default settings were used. The RSA tool 'Random Sequence Generator' was used to generate a random data set calibrated on non-coding upstream sequences of *Arabidopsis*. An *Arabidopsis*-specific Markov model was used to generate 10 random sequences. All random and *Eucalyptus* species sequences containing the mapped CREs were divided into 100-bp intervals and the CREs occurrences in each one were counted. For each 100-bp intervals of each promoter gene, a two-tailed *t*-test (assuming equal variance) was performed to identify those intervals, showing significant differences in the number of CREs relatively to random data set (for both, $\alpha = 0.01$ and $\alpha = 0.001$).

3.9 *Cis*-element conservation analysis

Cis-element conservation was assessed by manually counting the number and type of nucleotide changes that occurred within each *cis*-element occurrence considered. The *cis*-elements occurrences were classified as conserved, if no polymorphism occurred in any of the *Eucalyptus* species analyzed; putative *cis*-element occurrences that had a single nucleotide polymorphism or that had two or more nucleotide changes causing the loss of the respective *cis*-element in one or more species were also scored separately. All the occurrences absent in only one of the nine *Eucalyptus* species or absent in two or more species due to the same type of changes were annotated. The *cis*-element counts and the respective calculations of averages and percentages of the polymorphisms affecting *cis*-element occurrences were carried out using Excel (Microsoft Office 2007).

4 RESULTS

4.1 Extraction Integrity and Quantification of Eucalyptus species DNA

DNA was successful extracted from lyophilized leaves. A visible sharp band at the level of the λ-phage DNA band confirmed the presence of DNA. The average of overall DNA extracted from 62 samples was 5.072µg/µL, ranging from 0.196µg/µL (*E. amplifolia*, sample 54B3) to 1.030µg/µL (*E. globulus*, sample 47A1). The quantification using NanoDrop ND-1000 Spectrophotometer, confirm the purity of DNA, showing a 260/280 ratio of ~ 2.1. The 260/230 ratio had values of approximately 1.9, which are slightly lower than the desired values (2.00-2.20). However, the quantity and quality of the DNA obtained revealed viable for the present study.

4.2 Primer Testing for PCR Amplification

Twenty-three out of 30 (77%) primers sets yielded successful (a single band per locus) amplification for at least one of the species used in primer testing, according to the established PCR conditions (3.4 Material and Methods) and optimization. Only ProC3H3_PP primer set showed lack of amplification. The remaining six primer sets (20%) showed multiband amplification. The results of this prior test for PCR amplification were used to help establish the criteria for selection of the primer sets to further study. Therefore, a total of 11 primer sets (Table 1) were selected for amplification of the promoter regions over the 62 species (Table S1). The successful amplification over the eleven primer sets was very substantial (73%), showing a considerable degree of conservation of the promoter region in the related species under analysis (Table 1). Only Pro4CL1_PP showed a total percentage of successful amplification below 50% (42.6%). By contrast, ProCSE_PP (100%) led to a successful amplification in all the species analyzed.

Observing the results, it is possible to recognize that the species with greater failure in amplification, considering the results of the 11 primer sets, were the four species from the genus *Corymbia*. The species of the subgenus *Eucalyptus* showed successfully amplification values substantially lower than the observed in the subgenus *Symphyomyrtus*. Within the genus *Eucalyptus,* the species with lower successful amplification (less than 50%) were the ones belonging to the subgenus *Eucalyptus*, *E. elata*, *E. pilularis*, *E. piperita* and *E. stricta* (Table 1).

4.3 Isolation and analysis of promoter regions of the five lignin genes in nine *Eucalyptus* species and one *Corymbia* (*C. citriodora*).

Five primer sets were selected for amplification and further analysis of promoter regions: *EgrPAL9*, *EgrC4H2*, *EgrF5H2*, *EgrCAD3* and *EgrCSE*. The selection of these genes was done taking into account the position of the enzymes in the lignin biosynthetic pathway (Fig. 1). *EgrCAD3* gene (Eucgr.H03208) was removed from the current *E. grandis* v1.1 annotation but RNAseq and RT-qPCR expression data support that it is being actively transcribed in developing xylem tissues [13]. The gene can still be found in the low confidence transcripts in the current v1.1 annotation. Furthermore, these genes were selected because they could become of interest for future studies enlightening the integration of information obtained from different types of expressional regulation. *CSE* is a newly identified gene, recently described in *Arabidopsis* and not yet reported in any other species [16]. This gene was found to contribute for the lignification of plant vascular tissues, and also found to be highly and preferentially expressed in forming xylem in *Eucalyptus* [13].

Emphasis was given to the subgenus *Symphyomyrtus,* the largest and most diverse *Eucalyptus* subgenera [1], for species selection. Given their representativeness in terms of phylogenetic distribution, commercial importance and availability of plant samples, the following species were selected: *E. globulus* and *E. viminalis* (section *Maidenaria*), *E. grandis*, *E. urophylla, E. botryoides* (section *Transversaria*), *E. camaldulensis* and *E. tereticornis* (section *Exsertaria)* [4, 15, 60]. Furthermore, two distantly related species *E. regnans* and *E. elata*, from subgenus *Eucalyptus* (section *Renantheria*), were also selected for the analysis [4, 61]. Additionally, *C. citriodora* was also selected as an out group. Despite some disagreements [4], there is a general acceptance to consider *Corymbia* as different genus of eucalypts [3, 6], and several molecular studies support this classification [5, 62].

The 1000bp upstream ATG region of the five genes were amplified from one individual of each species and used to study patterns of sequence, spatial conservation and occurrence of CREs among *Eucalyptus* species, using PCR conditions described in Table 1. Even after several attempts of primer optimization and design it was not possible to isolate the upstream regions of the *PAL9*, *C4H2*, *F5H2* and *CAD3* from *E. elata* and *C. citriodora* may be due to high sequence divergence in the upstream priming sites. As a result, promoter data set relative to these four genes only contained sequences from eight *Eucalyptus* species.

Table 1: PCR conditions for amplification of 1000bp upstream ATG region of eleven lignin and success of amplification results by genera (*Corymbia vs Eucalyptus*) and *Eucalyptus* subgenera (*Eucalyptus vs Symphyomyrtus*)

Multi-gene family	**Gene short name**	**Primer ID**	**Ta (°C)**	**[MgCl2] mM**	**Phylogenetic divisions of species analyzed**											
					genus *Corymbia* (four species)			**genus *Eucalyptus* subgenus *Eucalyptus*** (13 species)			**genus *Eucalyptus* subgenus *Symphyomyrtus*** (43 species)			**Total of species** (62 species)		
					Amplification Features[a]			Amplification Features			Amplification Features			Amplification Features		
					A	**MA**	**NA**	**A**	**MA**	**NA**	**A**	**MA**	**NA**	**A**	**MA**	**NA**
Phenylalanine ammonia lyase (*PAL*)	*EgrPAL9*	ProPAL9_PP	60	1.5	25%	0%	75%	69.2%	61.5%	46.2%	81.8%	2.3%	2.3%	83.9%	1.6%	14.5%
Cinnamate 4- hydroxylase (*C4H*)	*EgrC4H1*	ProC4H1_PP	60	2.5	0%	25%	75%	0.0%	30.8%	38.5%	95.3%	4.7%	0.0%	82.0%	14.8%	3.3%
Cinnamate 4- hydroxylase (*C4H*)	*EgrC4H2*	ProC4H2_PP	65	1.5	0%	100%	0%	30.8%	7.7%	7.7%	72.1%	30.2%	0.0%	59.7%	38.7%	1.6%
Coumarate CoA ligase (*4CL*)	*Egr4CL1*	Pro4CL1_PP	65	1.5	0%	25%	75%	69.2%	61.5%	46.2%	51.2%	46.5%	2.3%	42.6%	47.6%	9.8%
Shikimate O-hydroxycinnamoyltransferase (*HCT*)	*EgrHCT4*	ProHCT4_PP	60	2.5	25%	50%	25%	0.0%	30.8%	38.5%	58.1%	41.9%	0.0%	60.7%	37.7%	1.6%
p-coumarate 3-hydroxylase (*C3H*)	*EgrC3H4*	ProC3H4_PP	60	2.5	75%	25%	0%	30.8%	7.7%	7.7%	72.1%	27.9%	0.0%	78.7%	21.3%	0.0%
Caffeoyl Shikimate Esterase (*CSE*)	*EgrCSE*	ProCSE_PP	60	2.5	100%	0%	0%	69.2%	61.5%	46.2%	100.0%	0.0%	0.0%	100.0%	0.0%	0.0%
Ferulate 5-hydroxylases (*F5H*)	*EgrF5H2*	ProF5H2_PP	60	2.5	0%	25%	75%	0.0%	30.8%	38.5%	95.5%	4.5%	0.0%	88.7%	4.8%	6.5%
Caffeic acid O-methyltransferase (*COMT*)	*EgrCOMT35*	ProCOMT35_PP	60	2.5	0%	0%	100%	30.8%	7.7%	7.7%	83.7%	7.0%	9.3%	70.5%	4.9%	24.6%
Cinnamyl CoA reductase (*CCR*)	*EgrCCR1*	ProCCR1_PP	60	2.5	0%	25%	75%	69.2%	61.5%	46.2%	93.0%	4.7%	2.3%	80.3%	6.6%	13.1%
Cynnamyl-alcohool dehydrogenase (*CAD*)	*EgrCAD3*	ProCCR3_PP	65	1.5	0%	0%	100%	0.0%	30.8%	38.5%	68.2%	22.7%	9.1%	56.5%	19.4%	24.2%
				Average	20%	25%	55%	34%	36%	33%	79%	17%	2%	73%	18%	9%

[a]**A** - *successful amplification*; **MA** – *multiband amplification*; **NA** – *no amplification.*

All the PCR products were initially, directly sequenced. Products that its sequences presented multiple overlapping sequence peaks (*CSE* - 95C7, 66B3, 6C4, 20B7, 8C5, 12C5 and 28A2; *F5H2* - 20B7 and 22B1) were discarded. For these cases corresponding PCR products were cloned before sequencing. In general, the efficiencies of cloning and transformation were found low (with few number of colonies and high number of false positives detected by PCR colony screening) (Fig. 2). Several attempts were carried out in order to increase the transformation efficiencies: volumes of PCR product, cloning reaction conditions, different competent *E.coli* cells strains and transformation reaction conditions.

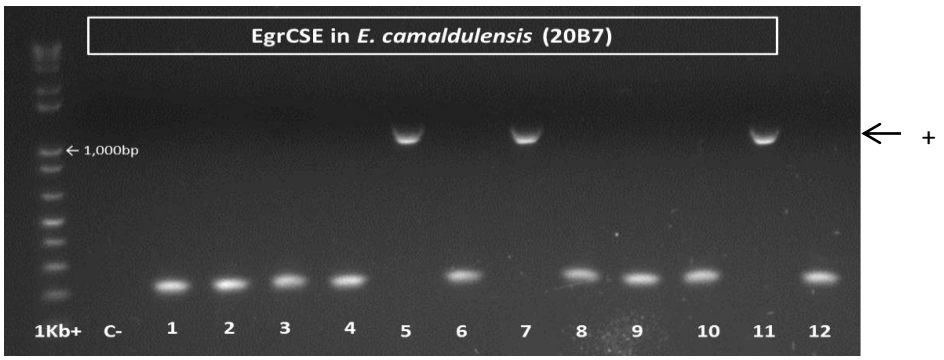

Fig. 2: Example of the results of PCR colony screening to *EgrCSE* in *E. camaldulensis.* The bands 5, 7 and 11 correspond to positively transformed clones with expected size – expected product size (951bp) and fragment of the vector amplified using universal (M13) vector primers (201bp).

For three of the species, the PCR products of *EgrC4H2* presented a non-specific band. Furthermore, PCR products successfully purified and sequenced always presented multiple overlapping peaks and poor quality along all promoter regions, and it would require the cloning of all PCR products for all species. Thus, for the purpose of this work and by time constraints it was decided to leave the study of the promoter region of this gene for further analysis.

To validate the correct sequence of all promoter sequences analyzed in this study, they were blasted against *E. grandis* genome [63]. The sequences matched a single region directly upstream of the corresponding gene, with high homology (> 88% nucleotide identity) (Table S3). This supports the inference that no paralogous promoters were isolated. After sequence treatment and trimming, promoter regions obtained varied in length from 421bp for *CAD3* in *E. urophylla* (6C4) to 992bp for the *PAL9* promoter in *E. botryoides* (Table S4). Some of the sequences obtained displayed significantly smaller size than the remaining placed in the same alignment. And in such cases their 3 'end aligned some base pairs upstream the ATG in the reference sequence.

4.4 Species-level DNA sequence variation and diversity in the promoter regions of *Eucalyptus* lignin genes.

Nucleotide diversity (*π,* [53]) and other measures of genetic polymorphism and sequence conservation were calculated using DnaSP version 5.10.01 software [52]. The distribution of nucleotide diversity along the promoter sequences of each gene was calculated using different subsets of sequences, due to the small size of some of the consensus sequences, except for *F5H2* promoter set (Table 2).

The mean G/C content was 43%, with a minimum of 33% in the *F5H2* promoter sequences. The highest G/C content was found in the promoter sequences of *PAL9*, with 53% (average of GC content values among the A, B and C subsets).

The average species-level nucleotide diversity (π) of the four *Eucalyptus* lignin genes promoter regions varied from π=0.026 for *PAL9* (average of A, B and C nucleotide diversity calculations) to π=0.078 for *CAD3* (average of A, B, C and D).

Table 2: Species-level nucleotide diversity in the promoter regions of four lignin genes from 9 *Eucalyptus* tree species

Promoter region ID	***PAL9***	***PAL9***	***PAL9***	***CSE***	***CSE***	***CSE***	***F5H2***	***CAD3***	***CAD3***	***CAD3***	***CAD3***
Species subsets[a]	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**A**	**B**	**C**	**D**
Number of species analyzed	8	7	6	9	8	10	8	8	7	7	5
Length of aligned sequence (including gaps/missing data) (bp)	920	920	920	886	886	924	795	904	904	904	904
G+C content (%)	55	50	54	46	45	46	33	31	37	34	38
Total number of sites (excluding gaps) (bp)	209	495	455	556	694	468	581	298	596	412	797
Number of polymorphic sites	57	56	54	112	99	190	86	155	150	148	112
Nucleotide diversity (π)	0.02428	0.02572	0.02664	0.03775	0.03603	0.05809	0.03172	0.07812	0.07770	0.07859	0.06971
Total number of insertions and deletions (indels) events analyzed	5	7	10	4	6	15	12	8	15	8	13

[a] **A (in the four promoter regions)** - all the *Eucalyptus* sequences available; **B (*PAL9*)** - without *E. botryoides* (12C5) sequence; **B (*CSE* and *CAD3*)** - without *E. urophylla* (6C4) sequence; **C (*PAL9*) -** without *E. urophylla* (6C4) and *E. regnans* (28A2) sequences; **C (*CSE*)** – including *C. citriodora* (8C5) sequence; **C (*CAD3*)** - without *E. camaldulensis* (20B7) sequence; **D (*CAD3*)** - without *E. urophylla* (6C4), *E. camaldulensis* (20B7) and *E. regnans* (28A2) sequences.

The graphs of nucleotide diversity distribution for each subset were overlapped to evaluate the distribution throughout the entire promoter region in analysis (Fig. 3). The profiles of the distribution of nucleotide diversity were coincident for all subsets of sequences, with the exception of *CSE*-C subset (green line - including *C. citriodora* sequence).

The levels of nucleotide diversity observed in promoter regions at the species-level are, in general, superior to 0.02. Nevertheless, there were regions in all promoters where species-level diversity was below 0.02 that is the threshold of nucleotide diversity expected for conserved coding regions [64]. This relative sequence conservation could indicate functional restrictions within those regions and may contain clusters of *cis*-regulatory elements.

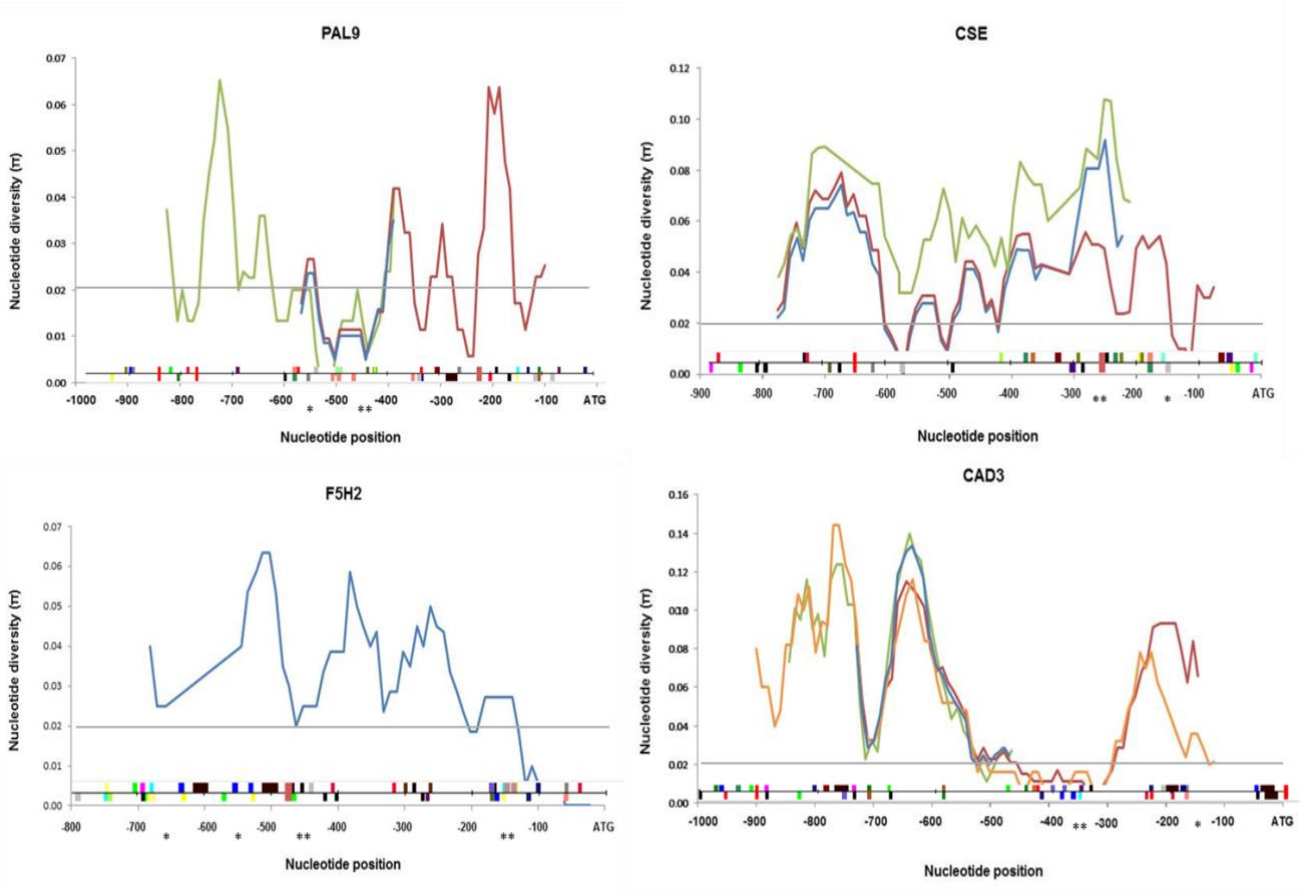

Fig. 3: Species-level nucleotide diversity profiles of the promoter regions from four *Eucalyptus* phenylpropanoid metabolism and lignin pathway genes in nine *Eucalyptus* and one *Corymbia* tree species. Nucleotide position is indicated relative to the start of translation (ATG position 0). Nucleotide diversity at 0.02 is marked by a grey line. The line and color blocks at the bottom of each graph show the position in the sense (upper) and anti-sense (down) strands of the mapped *cis*-element occurrences in the *E. grandis* reference sequence. A *cis*-element key is in Fig. 4 (See section 4.6) The different color lines correspond to the subsets of species listed in Table 2 as follows: blue – A; red – B; green – C and orange – D. Single asterisk and double asterisks at the bottom of each graph are positioned at the 100-bp intervals that showed higher number of *cis*-elements occurrences with significant differences from the random dataset. (p=0.01 and 0.001, respectively; two-tailed t-test assuming equal variance) (See section 4.6 and Fig.S2).

Indels and microsatellite regions were found in promoter regions of all the four genes. The *E. tereticornis PAL9* promoter region presented two 5bp indel (-488bp to -492bp and -635bp to -639bp) not observed in any of the other *Eucalyptus* species. In the *F5H2* promoter region a 103bp *indel* (-547bp to -650bp) was detected also for *E. tereticornis*. The *CSE* multiple alignment showed a 34bp conserved *indel* in the species *E. globulus*, *E. viminalis*, *E. camaldulensis*, *E. tereticornis* and *E. elata,* not being presented in the remaining species. It is interesting to point out to a microsatellite region (CT rich motif) between -145bp and -176bp in the promoter region of the CSE gene showing different polymorphisms in some of the species. In the *CAD3* promoter is also possible to observe a 17bp *indel* (-89bp to -105bp) in *E. grandis* and *E. globulus.*

4.5 *In silico* identification of putative CREs

4.5.1 Identification of known CREs

Cis-regulatory elements in the promoter sequences of genes members of the *Eucalyptus* core lignin *toolbox* and their orthologous in the other three genera were detected by homology in the plant-specific database PLACE. Excluding from the outset the motifs with five or less base pairs, a total of 219 different *cis*-elements were identified within the 75 promoter region sequences. These results were further refined to find CREs present in sequences belonging to the eleven lignin multigene families in *Eucalyptus*, which were identified in more than 50% of the sequences analyzed. Within this subset for the identified polyadenylation signals and functional TATA elements only the elements that appeared in the greater number of sequences were considered: TATABOX5 (89% of the sequences) and POLASIG1 (91% of the sequences). Table 3 contains detailed information of the selected 28 CREs from PLACE database who fulfilled those requirements. Those CREs are most likely involved in lignin gene regulation. Therefore, they were subsequently used to map the promoter regions of *Eucalyptus* species under study.

GT1CONSENSUS CRE was found in all the 75 gene promoters and MYCCONSENSUSAT and SEF4MOTIFGM7S occurred in 73. Twenty five percent of the CREs detected in those promoter regions were AC elements (MYB binding sites) (Table 3). The consensus sequence of some of those motifs could, potentially, lead to the recognition of the same sequence motif, resulting in the identification of more than one *cis*-element in the same position. For instance, BOXLCOREDCPAL (5'-ACCWWCC-3') could be identified in the same position than MYBPLANT (5'-MACCWAMC-3'). However, as there is no prior knowledge of which factors actually

bind to specific regions of the sequences in analysis, all the elements referred in Table 3 were used for *cis*-elements mapping. The five *cis*-regulatory elements retrieved from literature reports used to map the promoter regions of *Eucalyptus* species under study (3.7 Material and Methods) were also included in Table 3.

Table 3: Details of 33 *cis*-regulatory elements selected from PLACE database scans and literature and used for DNA pattern matching

Source[a]	Motif identity[b]	Motif sequence[c]	PLACE annotation[d]
Place DB	GT1CONSENSUS	GRWAAW	Light-regulated expression
	MYCCONSENSUSAT	CANNTG	Abiotic stress responsive element; Light and tissue-specific regulated expression
	SEF4MOTIFGM7S	RTTTTTR	Beta-conglycinin enhancer
	INRNTPSADB	YTCANTYY	Light response domain
	GT1GMSCAM4	GAAAAA	Pathogen- and salt-induced expression
	POLASIG1	AATAAA	Polyadenylation signal
	TATABOX5	TTATTT	Functional TATA element
	EECCRCAH1	GANTTNC	Enhancer element of CO2-responsive genes
	MYBPZM	CCWACC	phlobaphene pigmentation expression
	MYB1AT	WAACCA	Dehydration-responsive and ABA-responsive element
	MYBCORE	CNGTTR	Dehydration-responsive element
	BOXLCOREDCPAL	ACCWWCC	Elicitor treatment response element
	OSE1ROOTNODULE	AAAGAT	Organ-specific element
	PYRIMIDINEBOXOSRAMY1A	CCTTTT	Gibberellin-responsive element
	MYBPLANT	MACCWAMC	Phenylpropanoids specific expression
	MARTBOX	TTWTWTTWTT	Cell-fate decisions
	ANAERO1CONSENSUS	AAACAAA	Anaerobic responsive element
	DPBFCOREDCDC3	ACACNNG	ABA-responsive and embryo-specification element
	CARGCW8GAT	CWWWWWWWWG	Gibberellin metabolism specific expression
	CIACADIANLELHC	CAANNNNATC	Circadian-specific expression
	REALPHALGLHCB21	AACCAA	Phytochrome regulatory elements
	-300ELEMENT	TGHAAARK	Endosperm-specific expression
	MYB2CONSENSUSAT	YAACKG	Dehydration-responsive and ABA-responsive element
	PRECONSCRHSP70A	SCGAYNRNNNNNNNNNNNNNNHD	Plastid responsive element
	BOXIINTPATPB	ATAGAA	Plastid specific expression
	RHERPATEXPA7	KCACGW	Root hair-specific element
	CPBCSPOR	TATTAG	NADPH-Protochlorophyllide Oxidoreductase expression
	NTBBF1ARROLB	ACTTTA	Tissue-specific expression and auxin induction
Creux et al. (2008)	CRPE31	GNGNAGNG	Unknown
	CPRPE28	NNGCATGC	Iron deficiency-responsive element
	CPRPE26	TCCTGCYG	Unknown
	CPRPE25	RCYSTGCCC	Phloem-specific expression
Logemann et al. (1995)	PALBOXPPC	YTYYMMCMAMCMMC	Elicitor or light-regulated expression

[a]Original source of the *cis*-element; [b]Published name or identity of the *cis*-element; [c]Published consensus sequences for the *cis*-elements motifs with ambiguous bases represented as IUPAC codes where W=A/T, M=A/C, R=A/G, K=T/G, S=G/C, Y=C/T and N represent any of the four bases; [d]Putative function of the cis-elements as reported in literature or the PLACE database

4.5.2 Identification of unknown CREs using three software programs

Prediction of putative CREs was carried out by detection of over-represented motifs in 1 kb upstream promoter sequences of a set of genes highly or preferentially expressed in differentiating secondary xylem using three software programs (3.2 Material and Methods).

MotifSampler prediction allowed the identification of 50 motifs with Log-likelihood (LL) scores varying between 762 and 997. Some of the return motifs represented exactly the same consensus sequences. In the Oligo-analysis search, all the motifs displayed a number of occurrences considerably higher than the random expectation value, according to the background model. The significance values (sig= - $\log_{10}$ (e-value)), indicating the number of false-positives that would be expected by chance, varied between 0.4 and 31.72 (a significance higher than 0 would be expected, by chance alone, once per sequence set) [21]. The RSA tool pattern assembly automatically assembled the significant words, discovered by Oligo-analysis search. Clusters of aligned oligonucleotides were generated and the resultant best 20 consensus sequences (with greater lengths) were displayed. Since there was no information about the distribution of the motifs occurrences among the input sequences, MEME search were carried out in two different runs. The selection of the correct type of the motif distribution improves the sensitivity and quality of the motif search. The two runs differed in the distribution options, one considering zero or one occurrence of a single motif per sequence and the other considering any number of repetitions per sequence. The obtained motifs from MEME search are shown in Table S5.

Based on the outputs of the three software programs a total of 77 motif sequences were identified as overrepresented in the *Eucalyptus* and respective *A. thaliana, P. trichocarpa and V. vinifera* orthologous promoter regions. However, the output of those three programs included redundant and strong mutually overlapping motifs. The 77 motifs were introduced in STAMP software in order to detect motif similarity and redundancy within the outputs of motif-finders. The motifs of each defined group were manually aligned and a consensus sequence was generated. This allowed obtain a nonredundant set of 10 putative motifs (Table 4; Fig. S1). Five out of ten motifs were not annotated and not considered in further analysis: three of them consist in repeat motifs, being very likely microsatellite regions (AGGAGGAGGAGG; AGAAGAAGA and ATATATATATATAT), and the two other were extremely ambiguous (GGKNGKKGGN and GSCGGKBG).

Table 4: Over-represented motifs identified by MEME, Oligo-analysis (RSAT) and MotifSampler in the promoters of gene members belonging to the *Eucalyptus* core lignin *toolbox* and respective orthologous promoter sequences, of *A. thaliana*, *P. trichocarpa* and *V. vinifera*

Motif identity	Motif sequence[a]	PLACE *cis*-element[b]	E-value[c]	PLACE *cis*-element function and identity[d]	Software program
LRPE1	GGRGKWGGTGA	CCACCAACCCCC	6.45E-09	Vascular-specific expression (ACIIPVPAL2)	Oligo-analysis (RSAT); MEME; MotifSampler
LRPE2	GGNKGGNG	CTCCCAC	4.66E-07	Light-regulated expression (BOXCPSAS1)	MotifSampler
LRPE3	ACCTAAC	GTTAGGTT	1.28E-11	Tissue-specific activation of phenylpropanoid biosynthesis genes (MYB26PS)	Oligo-analysis (RSAT)
LRPE4	AGRGAGRG	TCTCTCTCT	3.06E-12	Enhancer of gene expression (CTRMCAMV35S)	Oligo-analysis (RSAT); MEME
LRPE5	TTTTCTTTT	AATAGAAAA	1.18E-08	Sucrose Responsive Element (SURE1STPAT21)	Oligo-analysis (RSAT); MEME

[a]Consensus sequences for the motifs detected by the software programs with ambiguous bases represented as IUPAC codes where W=A/T, M=A/C, R=A/G, K=T/G, B= C/G/T and N represent any of the four bases.
[b]*Cis*-element in the PLACE database which most closely resembles the motif identified in this study.
[c] The e-values result from the average of the e-values of all motifs of the group that gave rise to the consensus sequence. The e-value for STAMP is indicated by the false discovery ratio (FDR).
[d] PLACE cis-regulatory element putative function and the identity of the element as represented on the PLACE database.

Comparison of the remaining five motif sequences with the PLACE database allowed their putative annotation. The putative CRE LRPE1 was identified by the three software programs used. This putative CRE showed similarity to ACIIPVPAL2, an element required for vascular-specific gene expression and proved to be involved in complex patterns of tissue-specific expression of a PAL2 gene promoter [27]. The other four putative CRE were identified just by one or two of the software programs and exhibited similarity with elements involved in light response (LRPE2, [65]), phenylpropanoid biosynthesis (LRPE3, [66]), gene expression enhancement (LRPE4, [67]) and sucrose response (LRPE5, [68]). Although all the five putative CRE have a hit with a PLACE *cis*-element, for some of them the similarity is questionable and may represent novel elements that have yet to be functionally characterized (Fig. S1).

4.6 CREs mapping in the promoter region sequences

The presence and position conservation of CREs in the *PAL9*, *F5H2*, *CSE* and *CAD3* promoter regions were analyzed by mapping occurrences of the 38 previously identified putative CREs (33 from PLACE and literature scans – Table 3; 5 from motif *de novo* discovery search – Table 4) in the promoter sequences of 9 *Eucalyptus* species (Fig. 4). A large numbers of CREs were identified in both sense and antisense strands. It is easily noticeable that, in the four groups of promoters, there is a pattern of

conservation of the identified CREs within the *Eucalyptus* species. That pattern is not present in the orthologs of the genera *Arabidopsis*, *Populus* and *Vitis* (Fig. 4).

Three of the CREs (CPRPE25, CPRPE26 and CPRPE28; Table 3) could not be found in any of the *Eucalyptus* promoter sequences studied, and the first two (CPRPE25, CPRPE26), nor in the *Arabidopsis*, *Populus* and *Vitis* orthologs. Twenty two out of 38 CREs are absent in all the *Eucalyptus* species on at least one of the four orthologous groups of lignin genes. The remaining 14 CREs occur in the four groups of promoters. Nonetheless, 7 of them appeared with a single occurrence (appearing only in one of the *Eucalyptus* sequences) in at least one of the promoter groups. Thus, the CREs with occurrences in all the promoter orthologous groups and appearing with multiple occurrences in more than one sequence per promoter group were: GT1CONSENSUS, SEF4MOTIFGM7S, MYBPZM, PYRIMIDINEBOXOSRAMY1A, RHERPATEXPA7, LRPE2 and CRPE31 (Fig. 4). The putative CRE with the highest total number of occurrences in *Eucalyptus* sequences was GT1CONSENSUS with 246 occurrences. Five out these CREs appeared to be positionally conserved within two groups of promoters (Table 5).

Table 5: Groups of promoters where CREs appear to be positionally conserved and respective position of the occurrences

CRE	Promoters	Positions
GT1CONSENSUS	PAL9 / F5H2	-169bp / -172 and -165bp
	PAL9 / CSE	-730 and -720bp / -734 and -730bp
	F5H2 / CSE	-474 and -469bp / -498 and -493bp
SEF4MOTIFGM7S	PAL9 / CSE	-828 and -820bp / -837 and -798bp
	F5H2 / CAD3	-472 and -468bp / -496 and -477bp
	F5H2 / CAD3	-692 and -687bp / -700 and -681bp
MYBPZM	PAL9 / CSE	-586 and -580bp / -625 and -585bp
LRPE2	CSE / CAD3	-190 and -184bp / -193 and -176bp
CRPE31	PAL9 / CSE	-545 and -539bp / -580 and -541bp

In the four promoter orthologous groups it was possible to detect regions along the sequence that differ significantly in the number of *cis*-elements occurrences when compared to randomized dataset generated with a uniform distribution (Fig. S2). In the Figure 3, for each promoter, sequence was divided in intervals of 100bp, and the intervals with significant higher *cis*-elements number of occurrences, when compared with the randomized dataset, are indicated. The *cis*-element content of specific regions of the promoter sequences showing a nucleotide diversity particularly low (π below 0.02), that is indicating a potentially greater functional constraints, was analyzed in detail (Fig. 3).

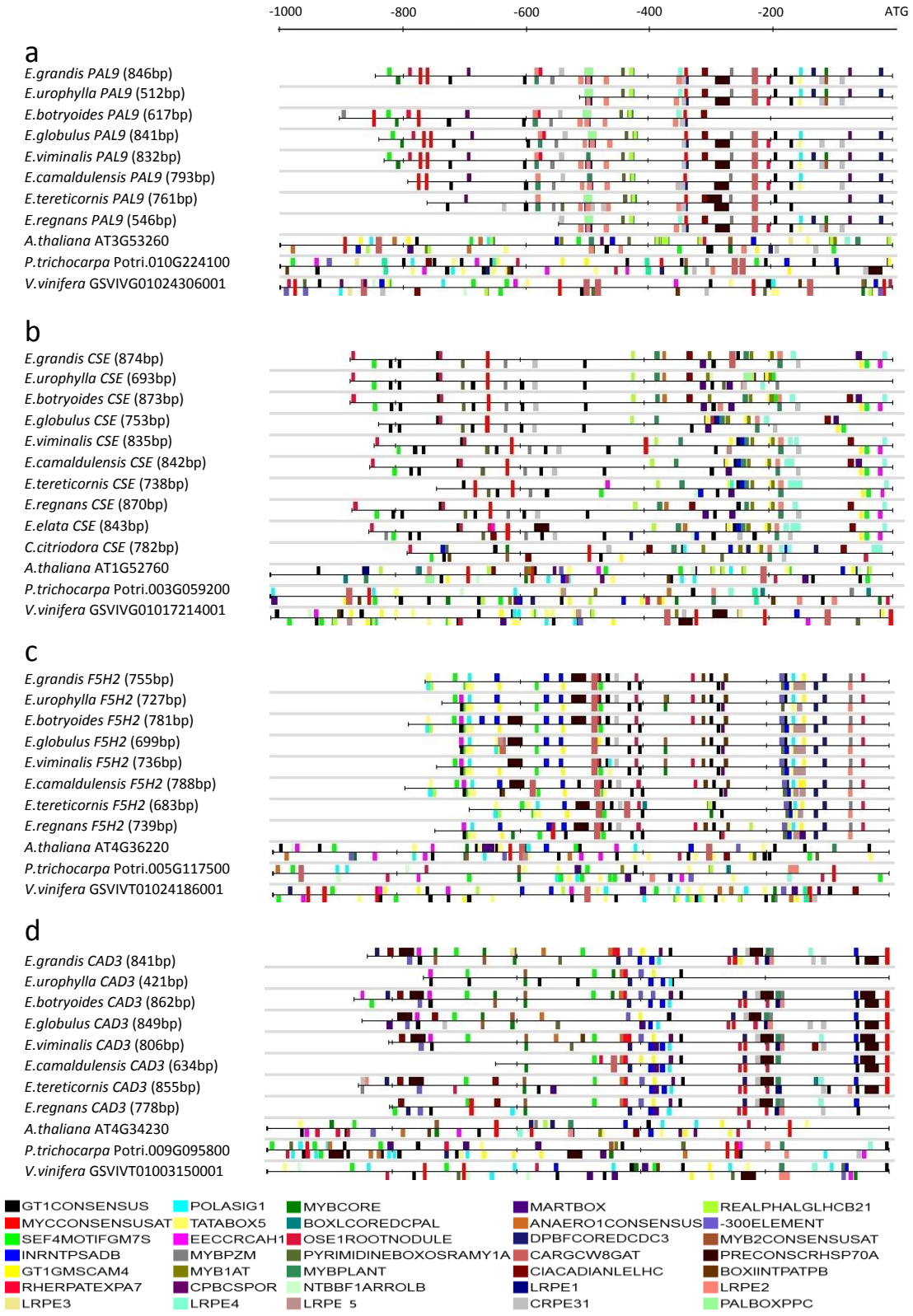

Fig. 4: Occurrences of 36 putative *cis*-regulatory elements mapped in the promoters of four orthologs groups of lignin genes in 9 *Eucalyptus* species, *C. corymbia*, *A. thaliana*, *P. trichocarpa* and *V. vinifera*. The size of each promoter region is indicated in brackets in *left-hand* margin along with name and species of each promoter (*PAL9* –a; *CSE* –b; *F5H2* –c and CAD3 –d). The relative positions of mapped *cis*-elements in relation to the translation start site (ATG) are indicated by the *ruler* at the top. A color *key* of *cis*-elements is given at the bottom of the image. *Horizontal black lines* in each group represent the promoter sequences for each species and *color blocks* show the position of the mapped *cis*-element occurrences found in sense (*above line*) or anti-sense orientation (*below line*).

In the cases where overlapping and redundant CREs were identified, was considered among them, the element who also occurred within a region of π below 0.02 in the other promoter regions (Table S6). All CREs with a probably conserved position between promoters of different genes (Table 5) were also identified in regions of particularly low nucleotide diversity (Fig. 3 and 4; Table S6), although not always in the same position. However, in some cases, the probably conserved positions between the promoters of two genes correspond to positions of low nucleotide diversity in those promoters.

4.7 *Cis*-element conservation analysis

To examine the potential effect of DNA sequence evolution on CRE occurrences in *Eucalyptus,* the putative CREs occurrences in the four lignin gene promoters founded in specific regions of particularly low nucleotide diversity (Fig. 3; Table S6) were analyzed for sequence conservation.

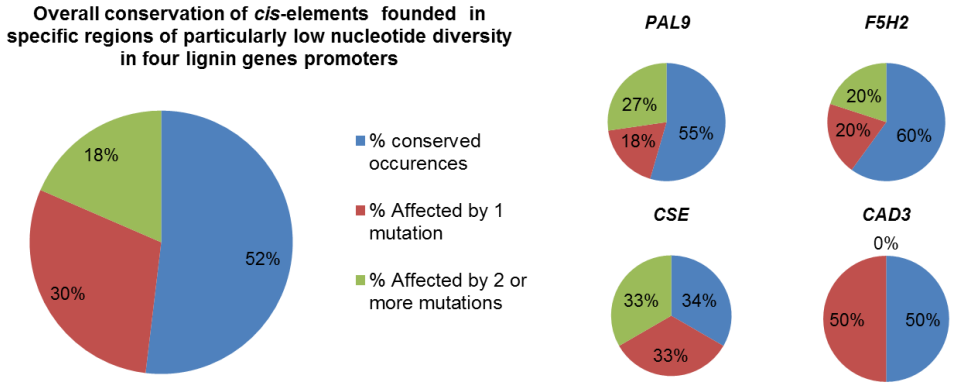

Fig. 5: Conservation of the *cis*-elements consensus sequences founded in specific regions of particularly low nucleotide diversity in four lignin gene promoters under study across studied *Eucalyptus* species

More than half of the *cis*-element occurrences (52%) distinguished in specific regions of low nucleotide diversity along the four promoter regions was totally conserved in the 9 *Eucalyptus* promoter sequences (Fig. 5). The *F5H2* promoters exhibited the highest number of fully conserved CREs occurrences (60%) and the *CSE* had the lowest (34%). Within the considered *cis*-element occurrences for conservation analysis there were no *cis*-element losses, in more than one species caused by the same type of nucleotide changes. On the other hand, nine cases (33% of the occurrences analyzed) were identified where a putative CRE occurrence was present in all but one of the nine *Eucalyptus* promoters, due to nucleotide changes in one of the species. Five out of the nine CREs losses (56%) occurred in *E. tereticornis*.

5 DISCUSSION AND CONCLUSION

Great economic implications in the control of wood quality for pulp and paper manufacturing have brought increasing interest in the knowledge of the *Eucalyptus* lignin biosynthesis genes. Transcriptional regulation of some of the genes, encoding several lignin biosynthetic enzymes is determinant in the control the lignification of secondary cell walls [12]. Despite several conserved *cis*-elements motifs have been discovered in the promoters of lignification genes questions still remain unanswered [11, 29]. In this work, the promoter regions of four phenylpropanoid-lignin metabolism genes across *Eucalyptus* tree species were characterized in terms of diversity, composition and conservation of *cis*-regulatory elements.

Amplification of promoter regions across 62 *Eucalyptus* species confirmed the existence of a considerable degree of conservation within the orthologous promoters (Table 1). The species of the same subgenus of *E. grandis*, for which the primers were developed, showed a remarkable success of amplification (79%). The increase of the evolutionary distance emphasizes the decreasing ability to successfully cross-species amplification due to the less preserved flanking sequences in species under study [69]. Thus, as expected, the species with less success in amplification were those belonging to the genus *Corymbia* (20%) and subsequently, those species belonging to the subgenus *Eucalyptus* (34%) (Table 1). Nevertheless, it is noteworthy that the CSE gene promoter region had 100% of successful amplification, suggesting that at least some parts of this promoter must be phylogenetically conserved among distant *Myrtaceae* species.

Based on the hypothesis that specific regions within promoters are more conserved than others [64] this work underwent a comparative genomic approach at the species-level [17]. The promoter regions of *PAL9*, *F5H2*, *CSE* and *CAD3* genes of nine *Eucalyptus* species (Fig. 4) were sequenced and aligned to get insight on the nucleotide diversity and for identification and mapping of CREs. In general the nucleotide diversity was high in all promoter regions. The inclusion of *C. citriodora* in the analysis had an important effect on nucleotide diversity over the analyzed region, as observed in *CSE* promoter graph (Fig. 3). The considerable lower conservation between these two orthologous sequences is not surprising since the divergence of the two clades, where these two genera belong date from early Paleogene [6]. The greater genetic diversity among populations of *E. tereticornis* compared with other *Eucalyptus* species was highlighted in several studies [18, 70]. It is worth noting that *E. tereticornis* presented *indels* not observed in any of the other species, in the promoters of *PAL9*

and *F5H2*. This is also reflected in the absence of several CREs only in this species. The differences in promoter region, in more than one gene, may have implications in transcriptional regulation of lignin genes in this species [71].

The nucleotide diversity values of the studied lignin genes promoter regions were high when compared to those observed in promoters of six cellulose synthase genes of different *Eucalyptus* species (minimum of 0.014 and maximum of 0.068) [17]. The *PAL9* promoter was the more conserved across the studied species with the lowest π value and the highest G/C content (Table 2). Rawal et al. (2012) investigated the conservation and divergence of the *PAL* gene family, observing a high percentage of orthology in all the varied groups of species represented [72]. Phenylalanine ammonialyase (*PAL*) catalyzes the first step of phenylpropanoid biosynthesis, being essential for the synthesis of all phenylpropanoids. Being essential may have caused a prominent maintenance of this promoter over divergence caused by speciation event [12]. In contrast, parameters of diversity in the *CAD3* promoter region, particularly the π, were substantially higher than in the other promoters. *CAD3* and *CAD2* genes, both members of the Cynnamyl-alcohool dehydrogenase (*CAD*) multigene family in *Eucalyptus,* were generated by recent segmental duplication [13]. The high levels of π observed might be explained by the simultaneous amplification of the promoter regions of both genes. However, the alignment of the *E. grandis CAD3* and *CAD2* promoter regions, revealed the existence of a 27bp *indel* that differentiate the two genes (-517 to -490bp). All the sequences obtained in this work match with *CAD3* within this specific region, being absent in *CAD2*, proving that all the sequences are *CAD3*. Still, the possibility of amplification of a gene duplication fragment or the amplification of a combination of sequences from the two different promoters [71] could be a possibility to explain the π levels observed for this promoter.

Despite the high average values of π, along the promoter, specific regions with nucleotide diversity values expected for conserved coding regions (π below 0.02) [17] were detected. This relative conservation may indicate functional restrictions within those regions.

The nature of the plant promoters and of the regulatory motifs makes difficult to distinguish conservation against the degenerate background frequency [22]. Most motif discovery algorithms report motifs although, often, they are just statistical artifacts. Thus, it becomes important evaluate the statistically significance of the motifs [23, 32]. Three out of the 5 novel putative CREs were identified by more than one motif discovery software program (Table 4).This is quite positive since the reliability and

accuracy in *cis*-element prediction is increased when it is found by different algorithms [15, 22, 23, 31]. Two novel putative CREs, LRPE1 and LRPE3, were similar to previous identified MYB protein binding sites (ACIIPVPAL2 and MYB26PS). MYB26 is a FT that recognizes *cis*-elements in the promoter regions of several phenylpropanoid biosynthetic genes [36, 66]. ACIIPVPAL2 is an element necessary for tissue specific expression of *PAL2* in different species, being also implicated in regulation of other lignin genes [27, 36]. The identification of motifs similar to known *cis*-elements of phenylpropanoid genes indicates a substantial robustness of the motif-finding algorithms used, giving more confidence to the results [40]. While all five motifs identified in this work showed similarity to previously identified elements in the PLACE database, the degree of similarity may be questioned in some cases (Table 4; Fig. S1). Thus, one cannot exclude the possibility that some of these motifs may be novel elements that have yet to be functionally characterized. However, none of the computational techniques for motif discovery can guarantee to find only biologically relevant motifs [23]. Despite all the efforts to choose the best search parameters there is also the possibility that none of the five elements detected (LRPE1 to 5) are functional CREs. This is because the used approach to detecting overrepresented motifs is extremely dependent on the specific dataset constructed and on the background model. This computational prediction allows studying novel binding sites that would normally not be identified by database searches and to ascertain if they are important in xylem forming specific genes regulation.

The search of CREs by homology allowed the identification of a set of motifs that appear to be significant and discriminatory, and abundantly distributed and duplicated, in the promoters of the co-regulated gene subset under analysis (Table 3). It is not surprising that a substantial part of the detected CREs in promoter regions were AC-rich, MYB-binding elements, responsible for specify vascular expression of phenylpropanoid genes in tissues in which lignification occurs [10, 11, 27, 29].

The mapping of the 38 putative CREs onto the promoter sequences of the selected genes showed that several of those MYB-binding elements were conserved in all the *Eucalyptus* promoters studied. For example, the BOXLCOREDCPAL [73] and MYBPLANT [74] were present in the promoters of all lignin genes analyzed, except in *F5H2* (Fig. 4). MYB58, a lignin specific TF in *Arabidopsis,* was proven to be able to induce the expression of the entire monolignol biosynthetic pathway genes, except *F5H,* by binding to AC elements [75]. It was proposed that, in *Arabidopsis*, *F5H* expression is directly regulated by the secondary cell wall master switch NST1/SND1, which is known to regulate expression of MYB58 [76]. The absence of these two

elements in *F5H2* promoters in *Eucalyptus* species could point to an activation of this gene similar to that which occurs in *Arabidopsis*.

The putative CREs CREP31, GT1CONSENSUS, SEF4MOTIFGM7S, PYRIMIDINEBOXOSRAMY1A, and LRPE2 were highly overrepresented, with multiple occurrences, in all the groups of promoters. The multiple occurrences of those CREs in the promoters may indicate that they are important in leveling de mRNA concentration or that they might play a role in up- or down-regulation of genes. This feature is characteristic when the synchronous expression of a subset of genes in the same tissue is observed [39]. Furthermore, all of them occur in regions with π below 0.02 in the promoter regions of one or more genes, having potentially greater functional constraints (Fig. 3 and 4; Table S6). In addition, all except PYRIMIDINEBOXOSRAMY1A seemed to be positionally conserved among the different genes (Table 5). This is relevant because the relative position of the regulatory elements have influence in the transcription profile of a gene [20, 71].

CRPE31 (Table 3, [15]) has no annotated function but was found in *Eucalyptus CesA* promoters in a *cis*-element cluster associated with the TSS. The reverse complement of this element could be an initiator element [17]. In *PAL9*, *CSE* and *F5H2* this element occurs in specific regions of low nucleotide diversity (Fig. 4; Table S6). Furthermore, it appeared to be positionally conserved in *PAL9* and *CSE* promoters (Table 5). These evidences support heavily its importance in the regulation of lignin biosynthesis.

The GT1CONSENSUS motif, the most overrepresented CRE, is a binding element of GT-1 proteins, which are conserved in plant nuclear genes and have diverse functions. This element, a putative target for the trihelix-related transcription factors, is annotated as being involved in plant responses to light and salicylic acid [75]. Rahantamalala et al. [43] had already showed the well conservation of GT-1 elements in the promoter regions of *CAD2* and *CCR* in various *Eucalyptus* species [43]. Those evidences point to a common control of lignin metabolism genes in response to light. The presence of the INRNTPSADB element in a particular conserved region of *CAD3* (-500 to -300bp) and the multiple occurrences of MYCCONSENSUSAT in *PAL9*, *CSE* and *CAD3* promoters are also consistent with light-regulated transcription of lignin genes [36]. They are related with light-dependent regulation in several genes namely of the phenylpropanoid biosynthesis (Table 3) [77, 78].

SEF4MOTIFGM7S element is the target binding site of SEF4, one of the four Soybean Embryo Factors which are related with differences in expression of the α' and ß subunit genes of beta-conglycinin seed storage protein [79]. Lessard et al. (1991) found that

the activity of this SEF is modulated over the course of seed development and speculate of its influence in spatial and temporally specific expression of the beta-conglycinin genes [79]. They mentioned that similar factors SEFs are common to many plant species. Thus, it is tempting to speculate that the occurrence of this element could imply the interaction of a TF similar to SEF4 important to modulate levels and timing of lignin genes expression.

LRPE2, the new putative CRE identified by motif discovery, was compared, by STAMP software, to the BOXCPSAS1 element in its reverse complement representation (Table 4; Fig. S1). In the PLACE database, this element is listed as a binding site involved in repression of the pea *AS1* gene (asparagine synthetase) induced by light. However the similarity of these two elements is questionable and instead LRPE2 may be a new CRE whose biological function needs further study.

PYRIMIDINEBOXOSRAMY1A was found in a region of low nucleotide diversity only in the *PAL9* promoter. This element is annotated in PLACE database as a gibbereling-responsive *cis*-element founded to be partially involved in the sugar repression of rice *α*-amylase gene (*RAmy1A)*, in rice embryos [80]. Nevertheless, Love et al. (2009) found that Gibberellic acid (GA) play a role in cambial cell differentiation and xylem development, as referred by Creux et al. (2013) [17]. Thus, this putative CRE may be involved in the transcriptional network regulating the secondary cell wall formation.

Logemann et al. (1995) prove that three CREs (boxes P, A, and L) present, alone or all of them together, in the promoters of *PAL* and *4CL* genes, in parsley, have functional relevance in gene activation mediated by elicitors or light [54]. The presence of BOXLCOREDCPAL (consensus of the putative sequences of box L) and PALBOXPPC (box P) in all *PAL9* sequences, in a specific region of particularly low nucleotide diversity, can be regarded as a clue to their involvement in elicitor or light mediated *PAL* gene activation in *Eucalyptus*.

The GT1GMSCAM4 element was detected in a specific region of π below 0.02 exclusively in *CAD3* promoter sequences. This element was found to play a role in pathogen induced gene expression of *SCaM-4* in soybean [81]. The presence of this element in *CAD3* promoter region may have a biological function since a number of *CAD* genes have been characterized for their response to plant pathogens [42].

Often *cis*-elements are dispersed into several distinct clusters [71]. In this work in most of the cases the significant higher number of CREs occurrences in a certain interval is due to the overlap of redundant elements. Therefore these results are biased to this

evaluation. Even so, it should be highlighted the -301bp to -400bp interval in *CAD3* promoter, once this is also a region of particular nucleotide diversity and where occur several non-redundant or overlapped elements. This region may be crucial in transcriptional control of *CAD* genes (Fig. 3; Table S6).

Sequence conservation in promoter region may not always indicate CREs conservation [31]. The analysis of the CREs content conservation of regions with low nucleotide diversity (Table S6) showed that more than half (52%) (Fig. 5) of occurrences were totally conserved in all *Eucalyptus* species. However, this value may be underestimated because those losses of a *cis*-element caused by one nucleotide change may also be conserved elements. CREs consensus sequences could have ambiguous bases that do not prevent the binding of the corresponding TF. In addition, some changes may be due to allelic variation within the different species. To clarify this aspect, further investigation at population level with a set of experimentally tested *cis*-elements, available for *Eucalyptus*, will be required. The high levels of conservation, points to the CREs maintenance in sequences, such as promoters, with fast evolution rate. The highest number of fully conserved CREs occurrences was exhibited in *F5H2* promoters. There are evidences supporting the crucial importance of protein stability and function of *F5H*, and thus it reduce tolerance to structural changes. The enzyme encoded by *F5H* has critical functions in the normal development of plants regulating lignification and lignin monomer composition [13].

In conclusion, this works provided new insight into the diversity and evolution of lignin genes promoters within *Eucalyptus*. Furthermore, sequences (CREs) that may be responsible for driving high or specific activity of promoters of genes of the lignin biosynthesis pathway in *Eucalyptus* were revealed. All putative CREs identified require experimental validation (e.g. insertion deletion experiments) [32, 71]. Nevertheless, the knowledge acquired, with the identification of novel motifs that are most likely to be real, offers a valuable foundation to accelerate the functional characterization of CREs. In turn, it will allow the identification of unknown TFs involved in the transcriptional control of lignin biosynthesis. For future work it will be interesting to make correlations between promoter regions with different polymorphisms and different composition and conservation of CREs with expression data. Additionally, the insertion-deletion polymorphisms founded among *Eucalyptus* species could potentially be used to specifically profile their genomes, by applying the PAAP-RAPD technique (promoter anchored amplified polymorphism based on random amplified polymorphic DNA) [69]. Combining all this knowledge will deepen the understanding of the mechanisms underlying the regulation of lignin heterogeneity in *Eucalyptus*.

References

1. Myburg, A.A., Grattapaglia, D., Tuskan, G.A., Hellsten, U., Hayes, R.D., Grimwood, J., Jenkins, J., Lindquist, E., Tice, H., Bauer, D. et al. (2014) The genome of Eucalyptus grandis. Nature, 510, 356-362.
2. Plomion, C., Leprovost, G. and Stokes, A. (2001) Wood formation in trees. Plant Physiol, 127, 1513-1523.
3. Boland, D., McDonald, M.W. and CSIRO. (2006) Forest trees of Australia. 5th ed. CSIRO Publishing, Collingwood, Vic.
4. Brooker, M.I.H. (2000) A new classification of genus Eucalyptus L'Hér. (Myrtaceae). Australian Systematic Botany, 13, 79–148.
5. Coppen, J.J.W. (2002) Eucalyptus: the genus eucalyptus. Taylor & Francis, London.
6. Bayly, M.J., Rigault, P., Spokevicius, A., Ladiges, P.Y., Ades, P.K., Anderson, C., Bossinger, G., Merchant, A., Udovicic, F., Woodrow, I.E. et al. (2013) Chloroplast genome analysis of Australian eucalypts--Eucalyptus, Corymbia, Angophora, Allosyncarpia and Stockwellia (Myrtaceae). Mol Phylogenet Evol, 69, 704-716.
7. FAO: Food and Agriculture Organization of the United Nations. (2005) State of the World's Forests 2005. 17 August 2014]; Available at: http://www.fao.org/docrep/008/a0400e/a0400e00.htm.
8. ICNF. (2013) IFN6 – Áreas dos usos do solo e das espécies florestais de Portugal continental. Resultados preliminares. [pdf], 34 pp, Instituto da Conservação da Natureza e das Florestas. Lisboa.
9. Rencoret, J., A. Gutierrez, and J. del Rio. (2007) Lipid and lignin composition of woods from different eucalypt species. Holzforschung, 61, 165–174.
10. Bonawitz, N.D. and Chapple, C. (2010) The genetics of lignin biosynthesis: connecting genotype to phenotype. Annu Rev Genet, 44, 337-363.
11. Bedon, F., Legay, S. (2011) CAB Reviews: Perspectives in Agriculture, Veterinary Science, Nutrition and Natural Resources, 6, 1.
12. Rogers, L.A. and Campbell, M.M. (2004) The genetic control of lignin deposition during plant growth and development. New Phytologist Volume 164, Issue 1. New Phytologist.
13. Carocha, V., Soler, M., Hefer C., Cassan-Wang, H., Myburg, A., Fevereiro, P., Paiva, J.A.P., Grima-Pettenati, J. Genome-wide analysis of the lignin toolbox of Eucalyptus grandis. (submitted).
14. Maleka, M.F. (2007) Allelic diversity in cellulose and lignin biosynthetic genes of Eucalyptus urophylla S. T. BLAKE. Pretoria: University of Pretoria. Master's dissertation.
15. Creux, N.M., Ranik, M., Berger, D.K. and Myburg, A.A. (2008) Comparative analysis of orthologous cellulose synthase promoters from Arabidopsis, Populus and Eucalyptus: evidence of conserved regulatory elements in angiosperms. New Phytol, 179, 722-737.
16. Vanholme, R., Cesarino, I., Rataj, K., Xiao, Y., Sundin, L., Goeminne, G., Kim, H., Cross, J., Morreel, K., Araujo, P. et al. (2013) Caffeoyl shikimate esterase (CSE) is an enzyme in the lignin biosynthetic pathway in Arabidopsis. Science, 341, 1103-1106.
17. Creux, N.M., De Castro, M.H., Ranik, M., Maleka, M.F., Myburg, A.A. (2013) Diversity and cis-element architecture of the promoter regions of cellulose synthase genes in Eucalyptus. Tree Gegetics & Genome, 9, 989-1004.
18. Külheim, C., Yeoh, S.H., Maintz, J., Foley, W.J. and Moran, G.F. (2009) Comparative SNP diversity among four Eucalyptus species for genes from secondary metabolite biosynthetic pathways. BMC Genomics, 10, 452.
19. Pereyra, N.B. (2009) In silico analysis of regulatory motifs in gene promoters. Barcelona: Universitat Pompeu Fabra. PhD thesis.
20. Sassaki, F.T. (2008) Isolamento e caracterização de promotores órgão-específicos a partir de informações do Banco FORESTs (Eucalyptus Genome Sequencing Project Consortium). São Paulo: Universidade Estadual Paulista, Instituto de Biociências de Botucatu. PhD thesis.
21. Thomas-Chollier, M., Defrance, M., Medina-Rivera, A., Sand, O., Herrmann, C., Thieffry, D. and van Helden, J. (2011) RSAT 2011: regulatory sequence analysis tools. Nucleic Acids Res, 39, W86-91.
22. Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J. et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol, 23, 137-144.

23. Bailey, T.L. (2008) Discovering sequence motifs. Methods Mol Biol, 452, 231-251.
24. Picot, E., Krusche, P., Tiskin, A., Carré, I. and Ott, S. (2010) Evolutionary analysis of regulatory sequences (EARS) in plants. Plant J, 64, 165-176.
25. Hauffe, K.D., Lee, S.P., Subramaniam, R. and Douglas, C.J. (1993) Combinatorial interactions between positive and negative cis-acting elements control spatial patterns of 4CL-1 expression in transgenic tobacco. Plant J, 4, 235-253.
26. Feuillet, C., Lauvergeat, V., Deswarte, C., Pilate, G., Boudet, A. and Grima-Pettenati, J. (1995) Tissue- and cell-specific expression of a cinnamyl alcohol dehydrogenase promoter in transgenic poplar plants. Plant Mol Biol, 27, 651-667.
27. Hatton, D., Sablowski, R., Yung, M.H., Smith, C., Schuch, W. and Bevan, M. (1995) Two classes of cis sequences contribute to tissue-specific expression of a PAL2 promoter in transgenic tobacco. Plant J, 7, 859-876.
28. Hatton, D., Smith, C. and Bevan, M. (1996) Tissue-specific expression of the PAL3 promoter requires the interaction of two conserved cis sequences. Plant Mol Biol, 31, 393-397.
29. Lacombe, E., Van Doorsselaere, J., Boerjan, W., Boudet, A.M. and Grima-Pettenati, J. (2000) Characterization of cis-elements required for vascular expression of the cinnamoyl CoA reductase gene and for protein-DNA complex formation. Plant J, 23, 663-676.
30. Goicoechea, M., Lacombe, E., Legay, S., Mihaljevic, S., Rech, P., Jauneau, A., Lapierre, C., Pollet, B., Verhaegen, D., Chaubet-Gigot, N. et al. (2005) EgMYB2, a new transcriptional activator from Eucalyptus xylem, regulates secondary cell wall formation and lignin biosynthesis. The Plant Journal Volume 43, Issue 4. The Plant Journal.
31. Reineke, A.R., Bornberg-Bauer, E. and Gu, J. (2011) Evolutionary divergence and limits of conserved non-coding sequence detection in plant genomes. Nucleic Acids Res, 39, 6029-6043.
32. Gertz, J., Riles, L., Turnbaugh, P., Ho, S.W. and Cohen, B.A. (2005) Discovery, validation, and genetic dissection of transcription factor binding sites by comparative and functional genomics. Genome Res, 15, 1145-1152.
33. Blanchette, M. and Tompa, M. (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting. Genome Res, 12, 739-748.
34. Bailey, T.L., Williams, N., Misleh, C. and Li, W.W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res, 34, W369-373.
35. Thijs, G., Marchal, K., Lescot, M., Rombauts, S., De Moor, B., Rouzé, P. and Moreau, Y. (2002) A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. J Comput Biol, 9, 447-464.
36. PLACE. A Database of Plant Cis-acting Regulatory DNA elements. May 2014]; Available at: http://www.dna.affrc.go.jp/PLACE/.
37. Higo, K., Ugawa, Y., Iwamoto, M. and Korenaga, T. (1999) Plant cis-acting regulatory DNA elements (PLACE) database: 1999. Nucleic Acids Res, 27, 297-300.
38. Vandepoele, K., Quimbaya, M., Casneuf, T., De Veylder, L. and Van de Peer, Y. (2009) Unraveling transcriptional control in Arabidopsis using cis-regulatory elements and coexpression networks. Plant Physiol, 150, 535-546.
39. Sharma, N., Russell, S.D., Bhalla, P.L. and Singh, M.B. (2011) Putative cis-regulatory elements in genes highly expressed in rice sperm cells. BMC Res Notes, 4, 319.
40. Shi, R., Sun, Y.H., Li, Q., Heber, S., Sederoff, R. and Chiang, V.L. (2010) Towards a systems approach for lignin biosynthesis in Populus trichocarpa: transcript abundance and specificity of the monolignol biosynthetic genes. Plant Cell Physiol, 51, 144-163.
41. Li, J., Yuan, J. and Li, M. (2014) Characterization of Putative cis-Regulatory Elements in Genes Preferentially Expressed in Arabidopsis Male Meiocytes. Biomed Res Int, 2014, 708364.
42. Raes, J., Rohde, A., Christensen, J.H., Van de Peer, Y. and Boerjan, W. (2003) Genome-wide characterization of the lignification toolbox in Arabidopsis. Plant Physiol, 133, 1051-1071.
43. Rahantamalala, A., Rech, P., Martinez, Y., Chaubet-Gigot, N., Grima-Pettenati, J. and Pacquit, V. (2010) Coordinated transcriptional regulation of two key genes in the lignin branch pathway--CAD and CCR--is mediated through MYB- binding sites. BMC Plant Biol, 10, 130.
44. Lauvergeat, V., Rech, P., Jauneau, A., Guez, C., Coutos-Thevenot, P. and Grima-Pettenati, J. (2002) The vascular expression pattern directed by the Eucalyptus gunnii

cinnamyl alcohol dehydrogenase EgCAD2 promoter is conserved among woody and herbaceous plant species. Plant Mol Biol, 50, 497-509.
45. Goes, E. (1985) Os Eucaliptos: identificação e monografia de 121 espécies existentes em Portugal. Lisboa: Portucel.
46. Gemas, V. (2004) Genetic variability of two woody perennials – Olea europaea L. and Eucalyptus globulus Labill. - assessed by RAPD and ISSR markers. Dissertation submitted to obtain a PhD in Biology.
47. PhytoExtract. Extraction of genomic sequences from Phytozome sources. October 2013]; Available at: http://www.polebio.lrsv.ups-tlse.fr/phytoExtract/.
48. BioMart. (2007) August 2014]; Available at: www.biomart.org
49. VecScreen. (2013) Screen a Sequence for Vector Contamination. April 2014]; Available at: http://www.ncbi.nlm.nih.gov/tools/vecscreen/.
50. Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res, 22, 4673-4680.
51. Hall, A. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser, 41, 95-98.
52. Librado, P., Rozas, J. (2009) DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics, 25, 1451-1452.
53. Nei, M., Li, W. (1979) Mathematical model for studying genetic variation in term of restriction endonucleases. Proc Natl Acad Sci USA, 76, 5267-5273.
54. Logemann, E., Parniske, M. and Hahlbrock, K. (1995) Modes of expression and common structural features of the complete phenylalanine ammonia-lyase gene family in parsley. Proc Natl Acad Sci U S A, 92, 5905-5909.
55. MotifSampler. (2014) August 2014]; Available at: http://bioinformatics.psb.ugent.be/webtools/MotifSuite/motifsampler.php.
56. RSAT: Regulatory Sequence Analysis Tools (1998) oligo-analysis: Analysis of oligomer occurrences in nucleotidic of peptidic sequences. August 2014]; Available at: http://rsat.ulb.ac.be/.
57. MEME: Multiple Em for Motif Elicitation. Version 4.9.1. August 2014]; Available at: http://meme.nbcr.net/meme/cgi-bin/meme.cgi.
58. STAMP: Alignment, Similarity, & Database Matching for DNA Motifs. September 2014]; Available at: http://www.benoslab.pitt.edu/stamp/.
59. Mahony, S. and Benos, P.V. (2007) STAMP: a web tool for exploring DNA-binding motif similarities. Nucleic Acids Res, 35, W253-258.
60. IPEF: Instituto de Pesquisas e Estudos Florestais (2005) Indicações para escolha de espécies de Eucalyptus. 17 Fevereiro 2014]; Available at: http://www.ipef.br/identificacao/eucalyptus/indicacoes.asp.
61. EucaLink: A Web Guide to the Eucalypts (2004). Eucalypt Classification. December 2013]; Available at: http://plantnet.rbgsyd.nsw.gov.au/PlantNet/Euc/class.html.
62. Parra-O., C., Bayly, M.J., Drinnan, A., Udovicic, F., Ladiges, P.Y. (2009) Phylogeny, major clades and infrageneric classification of Corymbia (Myrtaceae), based on nuclear ribosomal DNA and morphology. Australian Systematic Botany, 22, 384-399.
63. Phytozome v10. (2014) The JGI Comparative Plant Genomics Portal. May 2014]; Available at: http://phytozome.jgi.doe.gov/pz/portal.html.
64. Nei, M. (2007) The new mutation theory of phenotypic evolution. Proc Natl Acad Sci USA, 104, 12235-12242.
65. Ngai, N., Tsai, F.Y. and Coruzzi, G. (1997) Light-induced transcriptional repression of the pea AS1 gene: identification of cis-elements and transfactors. Plant J, 12, 1021-1034.
66. Uimari, A. and Strommer, J. (1997) Myb26: a MYB-like protein of pea flowers with affinity for promoters of phenylpropanoid genes. Plant J, 12, 1273-1284.
67. Pauli, S., Rothnie, H.M., Chen, G., He, X. and Hohn, T. (2004) The cauliflower mosaic virus 35S promoter extends into the transcribed region. J Virol, 78, 12120-12128.
68. Grierson, C., Du, J.S., de Torres Zabala, M., Beggs, K., Smith, C., Holdsworth, M. and Bevan, M. (1994) Separate cis sequences and trans factors direct metabolic and developmental regulation of a potato tuber storage protein gene. Plant J, 5, 815-826.
69. Poczai, P., Varga, I., Laos, M., Cseh, A., Bell, N., Valkonen, J.P. and Hyvönen, J. (2013) Advances in plant gene-targeted and functional markers: a review. Plant Methods, 9, 6.

70. Balasaravanan, T., Chezhian, P., Kamalakannan, R., Ghosh, M., Yasodha, R., Varghese, M. and Gurumurthi, K. (2005) Determination of inter- and intra-species genetic relationships among six Eucalyptus species based on inter-simple sequence repeats (ISSR). Tree Physiol, 25, 1295-1302.
71. Wray, G.A., Hahn, M.W., Abouheif, E., Balhoff, J.P., Pizer, M., Rockman, M.V. and Romano, L.A. (2003) The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol, 20, 1377-1419.
72. Rawal, H.C., Singh, N.K. and Sharma, T.R. (2013) Conservation, Divergence, and Genome-Wide Distribution of PAL and POX A Gene Families in Plants. Int J Genomics, 2013, 678969.
73. Maeda, K., Kimura, S., Demura, T., Takeda, J. and Ozeki, Y. (2005) DcMYB1 acts as a transcriptional activator of the carrot phenylalanine ammonia-lyase gene (DcPAL1) in response to elicitor treatment, UV-B irradiation and the dilution effect. Plant Mol Biol, 59, 739-752.
74. Sablowski, R.W., Moyano, E., Culianez-Macia, F.A., Schuch, W., Martin, C. and Bevan, M. (1994) A flower-specific Myb protein activates transcription of phenylpropanoid biosynthetic genes. EMBO J, 13, 128-137.
75. Zhou, J., Lee, C., Zhong, R. and Ye, Z.H. (2009) MYB58 and MYB63 are transcriptional activators of the lignin biosynthetic pathway during secondary cell wall formation in Arabidopsis. Plant Cell, 21, 248-266.
76. Zhao, Q., Wang, H., Yin, Y., Xu, Y., Chen, F. and Dixon, R.A. (2010) Syringyl lignin biosynthesis is directly regulated by a secondary cell wall master switch. Proc Natl Acad Sci U S A, 107, 14496-14501.
77. Hartmann, U., Sagasser, M., Mehrtens, F., Stracke, R. and Weisshaar, B. (2005) Differential combinatorial interactions of cis-acting elements recognized by R2R3-MYB, BZIP, and BHLH factors control light-responsive and tissue-specific activation of phenylpropanoid biosynthesis genes. Plant Mol Biol, 57, 155-171.
78. Nakamura, M., Tsunoda, T. and Obokata, J. (2002) Photosynthesis nuclear genes generally lack TATA-boxes: a tobacco photosystem I gene responds to light through an initiator. Plant J, 29, 1-10.
79. Lessard, P.A., Allen, R.D., Bernier, F., Crispino, J.D., Fujiwara, T. and Beachy, R.N. (1991) Multiple nuclear factors interact with upstream sequences of differentially regulated beta-conglycinin genes. Plant Mol Biol, 16, 397-413.
80. Morita, A., Umemura, T., Kuroyanagi, M., Futsuhara, Y., Perata, P. and Yamaguchi, J. (1998) Functional dissection of a sugar-repressed alpha-amylase gene (RAmy1 A) promoter in rice embryos. FEBS Lett, 423, 81-85.
81. Park, H.C., Kim, M.L., Kang, Y.H., Jeon, J.M., Yoo, J.H., Kim, M.C., Park, C.Y., Jeong, J.C., Moon, B.C., Lee, J.H. et al. (2004) Pathogen- and NaCl-induced expression of the SCaM-4 promoter is mediated in part by a GT-1 box that interacts with a GT-1-like transcription factor. Plant Physiol, 135, 2150-2161.

Supplementary material

Supplementary tables

Table S1: Classification and identification of *Eucalyptus* species samples from the collection plots in *Eucalyptus* Arboretum in National Forest in Escaroupim. Infrageneric classification follows Hill and Johnson (1995) for *Corymbia* and Brooker (2000) and EucaLink - A Web Guide to the Eucalypts (http://plantnet.rbgsyd.nsw.gov.au/PlantNet/Euc/class.html) within genus *Eucalyptus*.

Species name	Subgenus	Section [series]	Individual ID
Corymbia citriodora (Hook.) K.D. Hill & L.A.S. Johnson	-	[*Maculatae*]	8C5
Corymbia eximia (Schauer) K.D. Hill & L.A.S. Johnson	-	*Ochraria* [*Eximiae*]	10B3
Corymbia ficifolia (F. Muell.) K.D. Hill & L.A.S. Johnson	-	*Rufaria* [*Ficifoliae*]	103C2
Corymbia maculate (Hook.) K.D. Hill & L.A.S. Johnson	-	[*Maculatae*]	7A2
Eucalyptus albens Benth.	*Symphyomyrtus*	*Adnataria* [*Moluccanae*]	48A6
Eucalyptus amplifolia Naudin	*Symphyomyrtus*	*Exsertaria* [*Exsertae*]	54B3
Eucalyptus benthamii Maiden & Cambage	*Symphyomyrtus*	*Maidenaria* [*Benthamianae*]	71B7
Eucalyptus behriana F. Muell.	*Symphyomyrtus*	*Adnataria* [*Behrianae*]	138C2
Eucalyptus bicostata Maiden, Blakely & Simmonds	*Symphyomyrtus*	*Maidenaria* [*Globulares*]	53A1
Eucalyptus botryoides Sm	*Symphyomyrtus*	*Transversaria* [*Salignae*]	12C5
Eucalyptus calycogona Turcz.	*Symphyomyrtus*	*Bisectaria* [*Calycogonae*]	108C1
Eucalyptus camaldulensis Dehnh.	*Symphyomyrtus*	*Exsertaria* [*Exsertae*]	20B7
Eucalyptus cinerea F. Muell. ex Benth.	*Symphyomyrtus*	*Maidenaria* [*Cinereae*]	135B3
Eucalyptus cladocalyx F. Muell.	*Symphyomyrtus*	*Aenigmataria* [*Corynocalyces*]	121A5
Eucalyptus cneorifolia DC.	*Symphyomyrtus*	*Bisectaria* [*Cneorifoliae*]	58A5
Eucalyptus cornuta Labill.	*Symphyomyrtus*	*Bisectaria* [*Cornutae*]	18C1
Eucalyptus cosmophylla F. Muell.	*Symphyomyrtus*	*Triadaria* [*Longifoliae*]	61A2
Eucalyptus crebra F. Muell.	*Symphyomyrtus*	*Adnataria* [*Crebrae*]	2A6
Eucalyptus dalrympleana Maiden	*Symphyomyrtus*	*Maidenaria* [*Viminales*]	120C6
Eucalyptus dealbata A. Cunn. ex Schauer	*Symphyomyrtus*	*Exsertaria* [*Exsertae*]	101B3
Eucalyptus diversifolia Bonpl.	*Eucalyptus*	*Hesperia* [*Diversifoliae*]	139B3
Eucalyptus dives Schauer	*Eucalyptus*	*Renantheria* [*Amygdalinae*]	94B3
Eucalyptus drepanophylla F. Muell. ex Benth.	*Symphyomyrtus*	*Adnataria* [*Crebrae*]	68A2
Eucalyptus dumosa A. Cunn. ex Oxley	*Symphyomyrtus*	*Dumaria* [*Obtusiflorae*]	140A1
Eucalyptus elata Dehnh.	*Eucalyptus*	*Renantheria* [*Amygdalinae*]	95C7
Eucalyptus exserta F. Muell.	*Symphyomyrtus*	*Exsertaria* [*Exsertae*]	142B6
Eucalyptus globulus Labill.	*Symphyomyrtus*	*Maidenaria* [*Globulares*]	47A1
Eucalyptus gomphocephala DC.	*Symphyomyrtus*	*Bisectaria* [*Gomphocephalae*]	21C2
Eucalyptus goniocalyx F. Muell. ex Miq.	*Symphyomyrtus*	*Maidenaria* [*Globulares*]	57A7
Eucalyptus haemastoma Sm.	*Eucalyptus*	*Renantheria* [*Haemastomae*]	83A7
Eucalyptus macarthurii Deane & Maiden	*Symphyomyrtus*	*Maidenaria* [*Viminales*]	23A7
Eucalyptus macrorhyncha F.Muell. ex Benth	*Eucalyptus*	*Renantheria* [*Macrorhynchae*]	72A7
Eucalyptus maidenii F. Muell.	*Symphyomyrtus*	*Maidenaria* [*Globulares*]	107B6
Eucalyptus marginata Donn ex Sm.	*Symphyomyrtus*	*Bisectaria* [*Erythronemae*]	137C6
Eucalyptus melliodora A. Cunn. ex Schauer	*Symphyomyrtus*	*Adnataria* [*Melliodorae*]	70C3

Table S1: (continued)

Species name	Subgenus	Section [series]	Individual ID
Eucalyptus microcorys F. Muell.	*Nothocalyptus*	-	87B4
Eucalyptus microtheca F. Muell.	*Symphyomyrtus*	*Adnataria* [*Oliganthae*]	141B1
Eucalyptus moluccana Roxb.	*Symphyomyrtus*	*Adnataria* [*Moluccanae*]	76C2
Eucalyptus niphophila Maiden & Blakely	*Eucalyptus*	*Renantheria* [*Pauciflorae*]	67B5
Eucalyptus nitida Hook.	*Eucalyptus*	*Renantheria* [*Amygdalinae*]	134C2
Eucalyptus occidentalis Endl.	*Symphyomyrtus*	*Bisectaria* [*Occidentales*]	63A2
Eucalyptus ovata Labill.	*Symphyomyrtus*	*Maidenaria* [*Ovatae*]	74A4
Eucalyptus paniculata Sm.	*Symphyomyrtus*	*Adnataria* [*Paniculatae*]	11B1
Eucalyptus pauciflora Sieber ex Sprengel	*Eucalyptus*	*Renantheria* [*Pauciflorae*]	25A6
Eucalyptus pellita F. Muell.	*Symphyomyrtus*	*Transversaria* [*Resiniferae*]	88B5
Eucalyptus pilularis Sm.	*Eucalyptus*	*Renantheria* [*Pilulares*]	55A6
Eucalyptus piperita Sm.	*Eucalyptus*	*Renantheria* [*Piperitae*]	82C7
Eucalyptus polyanthemos Schauer	*Symphyomyrtus*	*Adnataria* [*Polyanthemae*]	1C3
Eucalyptus punctata DC.	*Symphyomyrtus*	*Transversaria* [*Punctatae*]	49C3
Eucalyptus regnans F. Muell.	*Eucalyptus*	*Renantheria* [*Obliquae*]	28A2
Eucalyptus resinifera Sm.	*Symphyomyrtus*	*Transversaria* [*Resiniferae*]	26C6
Eucalyptus robertsonii Blakely	*Eucalyptus*	*Renantheria* [*Amygdalinae*]	60A2
Eucalyptus robusta Sm.	*Symphyomyrtus*	*Transversaria* [*Salignae*]	14A5
Eucalyptus rubida Deane & Maiden	*Symphyomyrtus*	*Maidenaria* [*Viminales*]	102B6
Eucalyptus saligna Sm.	*Symphyomyrtus*	*Transversaria* [*Salignae*]	9B7
Eucalyptus salubris F. Muell.	*Symphyomyrtus*	*Bisectaria* [*Salubres*]	93C2
Eucalyptus stricta Sieber ex Sprengel	*Eucalyptus*	*Renantheria* [*Strictae*]	113C6
Eucalyptus tereticornis Sm.	*Symphyomyrtus*	*Exsertaria* [*Exsertae*]	22B1
Eucalyptus urophylla S.T. Blake	*Symphyomyrtus*	*Transversaria* [*Resiniferae*]	6C4
Eucalyptus viminalis Labill.	*Symphyomyrtus*	*Maidenaria* [*Viminales*]	66B3
Eucalyptus wandoo Blakely	*Symphyomyrtus*	*Bisectaria* [*Reduncae*]	117C4

Table S2: Novel primers used for end-to-end amplification of the *Eucalyptus* lignin genes *toolbox* promoter regions from *E. grandis* genomic DNA

Multi-gene family	Gene model	Gene short name	Scaffold	Expected Product size	Primer Name	PCR Primer Sequences (5'-3')
Phenylalanine ammonia lyase (PAL)	*Eucgr.A01144*	*EgrPAL1*	1	937bp	ProPAL1_Egr.A01144_PP	F: TGAGGTACAAAGATGAACAATCTGG R: TCAGCGACTTGGCCACCAA
Phenylalanine ammonia lyase (PAL)	*Eucgr.G02848*	*EgrPAL3*	7	979bp	ProPAL3_Egr.G02848_PP	F: TCAAAATCTCCACCCAGTCGG R: GACCATCCGCTTCACCTCAT
Phenylalanine ammonia lyase (PAL)	*Eucgr.G02849*	*EgrPAL4*	7	852bp	ProPAL4_Egr.G02849_PP	F: ACCCGACACATTTGTTAAGCA R: GGGACTCCTCGACCATCCTT
Phenylalanine ammonia lyase (PAL)	*Eucgr.G02850*	*EgrPAL5*	7	902bp	ProPAL5_Egr.G02850_PP	F: AGGGGTCACAATTTCATGGAT R: GGGACTCCTCGACCATCCTT
Phenylalanine ammonia lyase (PAL)	*Eucgr.G02851*	*EgrPAL6*	7	-	ProPAL6I_Egr.G02851_PP ProPAL6II_Egr.G02851_PP	F: AATACCATGGAGTCAGATAAGAT R: AGTCGACTGGTCAAAGCTGG; F: TAACCCGAGAGCCATCCTCA R: AGCAAGCGCGAGATCGAGGTGA

Multi-gene family	Gene model	Gene short name	Scaffold	Expected product size	Primer Name	PCR Primer Sequences (5'-3')
Phenylalanine ammonia lyase (PAL)	*Eucgr.G02852*	*EgrPAL7*	7	931bp	ProPAL7_Egr.G02852_PP	F: TGGTGATGAGCGACCAACAA R: CATCCGCTTCACCTCATCGA
Phenylalanine ammonia lyase (PAL)	*Eucgr.J01079*	*EgrPAL9*	10	985bp	ProPAL9_Egr.J01079_PP	F: TCACTCCTACCCACGGATGT R: AGTTCAGTGGGTCAGCATGG
Cinnamate 4-hydroxylase (C4H)	*Eucgr.C00065*	*EgrC4H1*	3	911bp	ProC4H1_Egr.C00065_PP	F: CGAGTGCCTCATGCTTAGCT R: AATGAGCGTGGATGCAAGGA
Cinnamate 4-hydroxylase (C4H)	*Eucgr.J01844*	*EgrC4H2*	10	752bp	ProC4H2_Egr.J01844_PP	F: CCCAAGTTGCAGTTGCCCTA R: GGAGCCAGTTGCCGAAGAT
Coumarate CoA ligase (4CL)	*Eucgr.C02284*	*Egr4CL1*	3	981bp	Pro4CL1_Egr.C02284_PP	F: ATGTGAGGTGAATGGCTGGG R: GTAGATGTCGGGGAGCTTCG
Shikimate O-hydroxycinnamoyltransferase (HCT)	*Eucgr.F03972*	*EgrHCT1*	6	1000bp	ProHCT1_Egr.F03972_PP	F: TGGCACCTTTAGCCCTCAAT R: GTGTGTCCTTAGCTGGGGAT
Shikimate O-hydroxycinnamoyltransferase (HCT)	*Eucgr.F03978*	*EgrHCT4*	6	931bp	ProHCT4_Egr.F03978_PP	F: AGAGTCCCAAAGCAAGGACG R: ACATTCGCATTCCACAGGGT
p-coumarate 3-hydroxylase (C3H)	*Eucgr.A02185*	*EgrC3H1*	1	999bp	ProC3H1_Egr.A02185_PP	F: CCAACCTTTGTCCCCCAGTT R: GAACCGCACGGACTTTGATG
p-coumarate 3-hydroxylase (C3H)	*Eucgr.A02188*	*EgrC3H2*	1	900bp	ProC3H2_Egr.A02188_PP	F: ACCCCTTTTTAGCGAAGGCA R: GGGGATGGAGAGGAGGATCA
p-coumarate 3-hydroxylase (C3H)	*Eucgr.A02190*	*EgrC3H3*	1	922bp	ProC3H3_Egr.A02190_PP	F: ACCCCACTTTTACACGCCAA R: CAGACCGATATGATGGGCCC
p-coumarate 3-hydroxylase (C3H)	*Eucgr.G03199*	*EgrC3H4*	7	939bp	ProC3H4_Egr.G03199_PP	F: TCCGACAAAGTCTCAACGGG R: TAACGCCGTAGAGGTTTCCG
Caffeoyl Shikimate Esterase (CSE)	*Eucgr.F02557*	*EgrCSE*	6	951bp	ProCSE_Egr.F02557_PP	F: GCAGCACGAGAAACATTGCA R: TGCGACGCGTAGTACTCATC
Cinnamoyl CoA O-methyltransferase (CCoAOMT)	*Eucgr.I01134*	*EgrCCoAOMT1*	9	905bp	ProCCoAOMT1_Egr.I01134_PP	F: GTCTGAACGACCAACCACCA R: AGAGACTTGTGGCCAACCTC
Cinnamoyl CoA O-methyltransferase (CCoAOMT)	*Eucgr.G01417*	*EgrCCoAOMT2*	7	981bp	ProCCoAOMT2_Egr.G01417_PP	F: TAATTGGATGCCGCCATGGA R: CAAGAACAACTGCCAAGGCC
Ferulate 5-hydroxylases (F5H)	*Eucgr.J02393*	*EgrF5H2*	10	986bp	ProF5H2_Egr.J02393_PP	F: TTCTCCGACGGGCATACAAG R: TCGCCCATCATGAGCATGTT
Caffeic acid O-methyltransferase (COMT)	*Eucgr.A01397*	*EgrCOMT1*	1	986bp	ProCOMT1_Egr.A01397_PP	F: ATTCAAGCCATCTGGACCGG R: CGGTCGAACCCATTCTCTCC
Caffeic acid O-methyltransferase (COMT)	*Eucgr.F02623*	*EgrCOMT35*	6	968bp	ProCOMT35_Egr.F02623_PP	F: TGCCACATGAGTTACGAAAAC R: AGGACCATTGGGAAGACGTG
Caffeic acid O-methyltransferase (COMT)	*Eucgr.K00951*	*EgrCOMT61*	11	963bp	ProCOMT61_Egr.K00951_PP	F: GCTTCACGAAGGACGACCAA R: CTTCCATATGTGCGCTTGGC
Caffeic acid O-methyltransferase (COMT)	*Eucgr.K00954*	*EgrCOMT63*	11	972bp	ProCOMT63_Egr.K00954_PP	F: GGGCCGAAGACGCAACTTAT R: TGTGTGCTTGGCCTCTTAGG
Cinnamyl CoA reductase (CCR)	*Eucgr.J03114*	*EgrCCR1*	10	937bp	ProCCR1_Egr.J03114_PP	F: ATAACGCTCTCGGCCGAAAA R: CGAAGGTGGCTAGCACTCAA
Cinnamyl CoA reductase (CCR)	*Eucgr.B02222*	*EgrCCR3*	2	961bp	ProCCR3_Egr.B02222_PP	F: TTTTCTCTTTGGCCGGTTGC R: GTTTACACGTGGGCCTACGA
Cynnamyl-alcohool dehydrogenase (CAD)	*Eucgr.G01350*	*EgrCAD2*	7	948bp	ProCAD2_Egr.G01350_PP	F: TACCCCTTCGACACATTGGC R: AACTCGAAACCCTGCCTGAG

Multi-gene family	Gene model	Gene short name	Scaffold	Expected product size	Primer Name	PCR Primer Sequences (5'-3')
Cynnamyl-alcohool dehydrogenase (CAD)	*Eucgr.H03208*	*EgrCAD3*	8	956bp	ProCAD3_Egr.H03208_PP	F: GAGCCACACGTACCTCTTCG R: AAGTGTAAGGCGCGAGAGTC
Cynnamyl-alcohool dehydrogenase (CAD)	*Eucgr.E01107*	*EgrCAD17*	5	935bp	ProCAD17_Egr.E01107_PP	F: GAGAGATGGACACGCCTTCG R: TCTCGCATGAACACTTCGGT
Phenylalanine ammonia lyase (PAL)	*Eucgr.J01079*	*EgrPAL9*	10	998bp 943bp	PAL9_A_Egr.J01079_PP PAL9_B_Egr.J01079_PP	F: TGTGGCAGTTGGTTGGGTTA R: TGAAAGCCGAGATTAGCCGT F: GTGCACGTATCGCAAGTTTCA R: TACTCCTCGACCATCCGCTT
Cinnamate 4-hydroxylase (C4H)	*Eucgr.J01844*	*EgrC4H2*	10	1270bp 893bp	C4H2_A_Egr.J01844_PP C4H2_B_Egr.J01844_PP	F: TCGCATCACATCATGTACGT R: GAGGGTCTTCTCCAGGAGGA F: AGAGAGTTACACCCAAGTTGCA R: AGCACCTCCTTGGAGAGGTC
Cynnamyl-alcohool dehydrogenase (CAD)	*Eucgr.H03208*	*EgrCAD3*	8	966bp 995bp	CAD3_A_Egr.H03208_PP CAD3_B_Egr.H03208_PP	F: GGTCCCAACAACTTTATGCACA R: TCACACTCGACGTGATTTGT F: AGGCGCCAAAACGACAAAAA R: CAGGCAAACAACCGCAGATC

Table S3: Results of the matching of promoter sequences, in this study, with the *E.grandis* genome sequences

Specie	PAL9					CSE				
	Sequence size (bp)	Scaffold	Score (bits)	E-value	Identity	Sequence size (bp)	Scaffold	Score (bits)	E-value	Identity
E. grandis	846	10	1523.3	0	98.50%	874	Eucalyptus grandis - JGI v1.1 sequence			
E. urophylla	512	10	967.9	0	99.60%	693	6	1144.6	0	96.70%
E. botryoides	992	10	1725.3	0	98.6%	873	6	1644.2	0	98.50%
E. globulus	841	10	1431.4	0	96.90%	753	6	1148.2	0	92.00%
E. viminalis	832	10	1537.8	0	98.3%	835	6	1404.3	0	92.40%
E. camaldulensis	793	10	1285.3	0	95.90%	842	6	1337.6	0	92.90%
E. tereticornis	761	10	1256.4	0	92.20%	738	6	1016.6	0	88.50%
E. regnans	546	10	1014.8	0	99.0%	870	6	1647.8	0	98.60%
E. elata	-	-	-	-	-	843	6	1353.8	0	91.00%
C. citriodora	-	-	-	-	-	782	6	1647.8	0	98.60%
Specie	**F5H2**					**CAD3**				
	Sequence size (bp)	**Scaffold**	**Score (bits)**	**E-value**	**Identity**	**Sequence size (bp)**	**Scaffold**	**Score (bits)**	**E-value**	**Identity**
E. grandis	755	10	1563	0	98.40%	841	8	746.1	0	94.10%
E. urophylla	727	10	1563	0	98.00%	421	8	226.7	1.5E-57	98.50%
E. botryoides	781	10	1627.9	0	99.50%	862	8	830.8	0	100.00%
E. globulus	699	10	1435	0	98.00%	849	8	692	0	89.80%
E. viminalis	736	10	1608.1	0	98.70%	806	8	782	0	97.80%
E. camaldulensis	788	10	1370	0	96.00%	634	8	764.1	0	97.00%
E. tereticornis	683	10	1164.5	0	93.90%	855	8	719	0	94.80%
E. regnans	739	10	1472.8	0	97.10%	778	8	675	0	94.90%
E. elata	-	-	-	-	-	-	-	-	-	-
C. citriodora	-	-	-	-	-	-	-	-	-	-

Table S4: Results of the matching of orthologs promoter sequences of *Arabidopsis*, *Populus* and *Vitis* using *Eucalyptus grandis* genes as target

Multi-gene family	Gene short name	Gene model	Corresponding Orthologs								
			Arabidopsis	**Score**	**Similarity**	***Populus***	**Score**	**Similarity**	***Vitis***	**Score**	**Similarity**
Phenylalanine ammonia lyase (***PAL***)	*EgrPAL1*	Eucgr.A01144	AT3G53260	3246	84.0%	Potri.010G224100	3370	85.6%	GSVIVG01024306001	2393	64.5%
	EgrPAL3	Eucgr.G02848		3529	86.8%		3701	90.5%	GSVIVT01024306001	2679	68.9%
	EgrPAL4	Eucgr.G02849		3471	85.4%		3605	88.4%		2538	65.5%
	EgrPAL5	Eucgr.G02850		3464	85.4%		3592	88.2%		2523	65.3%
	EgrPAL6	Eucgr.G02851		3564	88.1%		3726	91.2%		2669	68.9%
	EgrPAL7	Eucgr.G02852		3381	83.3%		3529	86.7%	GSVIVG01024292001	2526	66.7%
	EgrPAL9	Eucgr.J01079		3552	87.8%		3686	91.3%	GSVIVG01024306001	2624	68.5%
Cinnamate 4- hydroxylase (***C4H***)	*EgrC4H1*	Eucgr.C00065	AT2G30490	1731	64.7%	Potri.018G146100	2415	77.4%	GSVIVG01035166001	1542	53.4%
	EgrC4H2	Eucgr.J01844		2395	84.6%	Potri.013G157900	2638	89.1%	GSVIVG01024554001	2180	79.6%
Coumarate CoA ligase (***4CL***)	*Egr4CL1*	Eucgr.C02284	AT3G21240	2257	81.1%	Potri.006G169700	2549	87.3%	GSVIVG01029182001	2369	80.7%
Shikimate O-hydroxycinnamoyltransferase (***HCT***)	*EgrHCT1*	Eucgr.F03972	AT5G48930	1976	82.7%	Potri.003G183900	2093	85.2%	GSVIVT01016053001	1899	80.0%
	EgrHCT4	Eucgr.F03978		2085	82.8%		2234	85.7%	GSVIVG01016053001	2013	80.4%
	EgrHCT5	Eucgr.J03126		2264	78.5%		2337	81.1%		2130	74.2%
p-coumarate 3-hydroxylase (***C3H***)	*EgrC3H1*	Eucgr.A02185	AT2G40890	2545	85.1%	Potri.006G033300	2701	88.2%	GSVIVT01025800001	1368	47.8%
	EgrC3H2	Eucgr.A02188		2516	84.1%		2699	87.4%		1396	47.9%
	EgrC3H3	Eucgr.A02190		2683	89.2%		2898	93.1%		1532	51.7%
	EgrC3H4	Eucgr.G03199		2418	82.5%		2671	88.4%		1390	47.8%
Caffeoyl Shikimate Esterase (***CSE***)	*EgrCSE*	Eucgr.F02557	AT1G52760	1675	88.3%	Potri.003G059200	1813	91.4%	GSVIVG01017214001	1316	72.8%
Cinnamoyl CoA O-methyltransferase (***CCoAOMT***)	*EgrCCoAOMT1*	Eucgr.I01134	AT4G34050	1336	94.7%	Potri.009G099800	1429	98.0%	GSVIVT01022100001	1412	97.2%
	EgrCCoAOMT2	Eucgr.G01417		1367	94.7%	Potri.001G304800	1395	95.5%		1401	95.5%
Ferulate 5-hydroxylases (***F5H***)	*EgrF5H2*	Eucgr.J02393	AT4G36220	2218	78.1%	Potri.005G117500	2188	78.8%	GSVIVT01024186001	1952	70.7%
Caffeic acid O-methyltransferase (***COMT***)	*EgrCOMT1*	Eucgr.A01397	AT5G54160	1728	73.4%	Potri.012G006400	1912	76.0%	GSVIVT01008854001	1336	60.1%
	EgrCOMT35	Eucgr.F02623		1018	65.1%		1082	66.5%	GSVIVT01034498001	1305	71.9%
	EgrCOMT61	Eucgr.K00951		780	56.9%	Potri.019G102900	1346	74.4%	GSVIVT01020642001	808	58.6%
	EgrCOMT63	Eucgr.K00954		695	59.6%		1197	76.1%			
Cinnamyl CoA reductase (***CCR***)	*EgrCCR1*	Eucgr.J03114	AT1G15950	1192	49.9%	Potri.001G046100	1356	56.3%	GSVIVT01034241001	1400	57.0%
	EgrCCR3	Eucgr.B02222	AT5G58490	1456	71.3%	Potri.009G076300	1622	82.0%	GSVIVT01024672001	1506	72.4%
Cynnamyl-alcohool dehydrogenase (***CAD***)	*EgrCAD2/CAD3*	Eucgr.G01350	AT4G34230	1356	79.0%	Potri.009G095800	1438	80.8%	GSVIVT01003150001	1388	82.6%
	EgrCAD17	Eucgr.E01107	AT1G72680	1426	85.8%	Potri.011G148100	1526	89.2%	GSVIVT01019711001	1465	87.0%

Table S5: MEME motif search results

Type of distribution	Motif sequence	E value	Sites
Zero or one per sequence	CCA[AC]CC[CA]C	3.1E-23	49
	CACCAAC[CT]	1.5E-08	31
	AG[GA][GA]AG[AG]G	1.6E-03	48
Any number of repetitions	CACC[ATC]ACC	1.1E-36	50
	AG[AG]GAG[AG]G	5.4E-16	47
	CC[AT][AC]C[TC]CC	5.2E-11	50
	AAAGAAAA	1.1E-03	50

Table S6: CREs identified in regions of promoters with π below 0.02

CRE	Promoter region	Regions with π below 0.02
GT1CONSENSUS	CSE F5H2 CAD3	-500bp -330 to -300bp; -500 to -430bp -500 to -300bp
CRPE31	PAL9 CSE F5H2	-340bp; -160bp to -120bp -575bp -500 to -430bp
DPBFCOREDCDC3	PAL9 F5H2 CAD3	-340bp; -160bp to -120bp -120bp -500 to -300bp
SEF4MOTIFGM7S	PAL9 F5H2 CAD3	-800bp -500 to -430bp -500 to -300bp; -700bp
REALPHALGLHCB21	PAL9 CSE	-460bp -420bp
RHERPATEXPA7	F5H2 CAD3	-330 to -300Bbp -500 to -300bp
POLASIG1	PAL9 CAD3	-160bp to -120bp -500 to -300bp
CARGCW8GAT	F5H2	-500 to -430bp
LRPE2	PAL9	-500bp
LRPE4	PAL9	-340bp
PYRIMIDINEBOXOSRAMY1A	PAL9	-460bp
BOXLCOREDCPAL	PAL9	-500bp
LRPE1	PAL9	-500bp
PALBOXPPC	PAL9	-500bp
MYBCORE	PAL9	-800bp
CPBCSPOR	CAD3	-500 to -300bp
INRNTPSADB	CAD3	-500 to -300bp
ANAERO1CONSENSUS	CAD3	-500 to -300bp
GT1GMSCAM4	CAD3	-500 to -300bp

Supplementary Figures

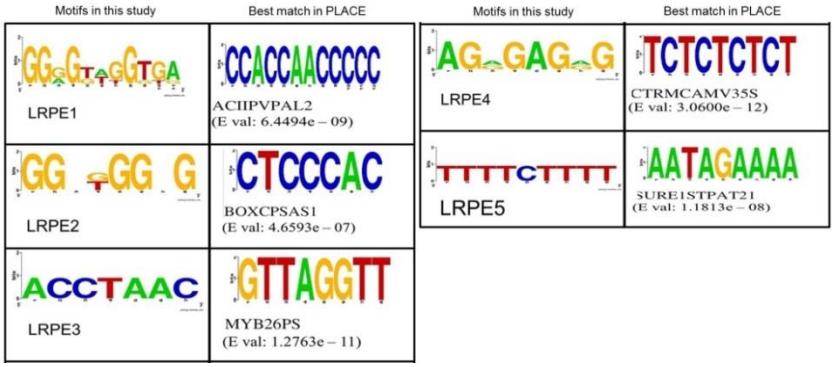

Fig. S1: Sequence logos of overrepresented sequences in the promoters of genes highly or preferentially expressed in secondary xylem and involved in phenylpropanoid-lignin metabolism in *Eucalyptus,* detected using MEME, MotifSampler and oligo-analysis (RSAT). Letters in the logos abbreviate the nucleotides (A, C, G and T) and are sized relative to their occurrence.

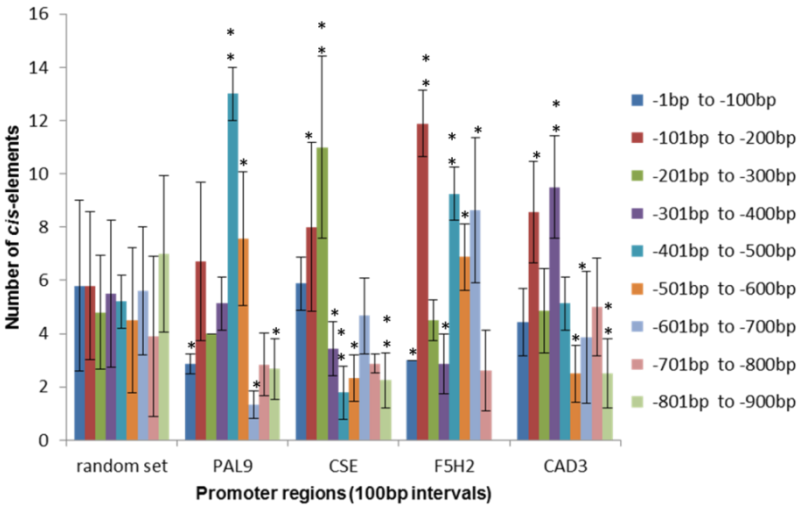

Fig. S2: Frequency of cis-elements occurrences along the length of the four lignin genes promoter regions averaged across the 9 *Eucalyptus* species in 100-bp intervals compared to a randomly generated sequence dataset. The *y-axis* gives the number of *cis*-elements in each 100-bp interval, and the *x-axis* represents the promoter regions being analyzed. The error bars show the standard deviation. Single *asterisk* and *double asterisks* indicate significant differences from the random dataset (p=0.01 and 0.001, respectively; two-tailed t-test assuming equal variance).